EVOLUTION DER SUBSTRATERKENNUNGSDOMÄNEN VON AAA PROTEINEN

EVOLUTION OF SUBSTRATE RECOGNITION DOMAINS OF THE AAA PROTEINS

DISSERTATION

der Fakultät für Chemie und Pharmazie

der Eberhard-Karls-Universität Tübingen

zur Erlangung des Grades eines Doktors

der Naturwissenschaften

2007

vorgelegt von

SERGEJ DJURANOVIC

Tag der mündlichen Prüfung: 09.11.2006

Dekan: Prof. Dr. Lars Wesemann

1. Berichterstatter: Prof. Dr. Gabriele Dodt
2. Berichterstatter: Prof. Dr. Andrei Lupas
3. Berichterstatter: Prof. Dr. Kai-Uwe Fröhlich

Table of contents

Evolution of substrate recognition domains

of the AAA proteins

1. Introduction

1.1 AAA proteins

The AAA proteins, a new family of „**A**TPases **A**ssociated with diverse cellular **A**ctivities" were first described by Erdmann et al. (1991). They found that proteins associated with different biological processes ranging from DNA repair and replication to organelle biogenesis, membrane trafficking, transcriptional regulation, and protein quality control, share a highly conserved domain responsible for ATP binding. The conserved ATPase domain (AAA) with a length of 200-250-amino acid residues was found to contain Walker A and B motifs - two sequences typical for P-loop nucleoside triphosphatases (NTPases) (Walker et al, 1982). These motifs are involved in binding of the triphosphate moiety of the substrate and coordination of an Mg^{2+} ion, which is important for subsequent hydrolysis of the ATP. Another region of high sequence conservation, the so-called 'second region of homology' (SRH) found between the two Walker motifs (Swaffield et al., 1992), makes AAA family distinguishable from the larger and more diverse family of AAA+ proteins (Lupas and Martin, 2002). The domain architecture of AAA proteins consists of a non-ATPase N-terminal domain, which is the putative substrate binding site, followed by one or two copies of AAA domains (named D1 and D2) (Fig. 1.1.). All AAA proteins whose oligomeric structure has been investigated up to now form hexameric rings (Fig. 1.2) (Hartman and Vale, 1999). Some AAA proteins, like katanin, can exist as dimers and only hexamerize in a substrate-dependent manner (Scott et al., 2005). Further oligomerization to dodecameric complexes was also noticed for some proteins (Wolf et al., 1998; Scott et al., 2005). In members that contain two copies of AAA domains, one of the domains may be degenerate such as D1 in p97/CDC48 or D2 in Sec18/NSF, and may be primarily involved in structural stability of the hexamer complex (Singh et al., 1999).
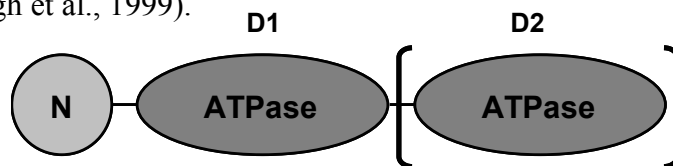
Figure 1.1 Schematic diagram of domain organization of the AAA proteins

There are several crystal structures of the AAA proteins or their AAA modules in the PDB database (http://www.rcsb.org/pdb/) that confirm their remarkably well conserved structure and mechanism of interaction with nucleotides.
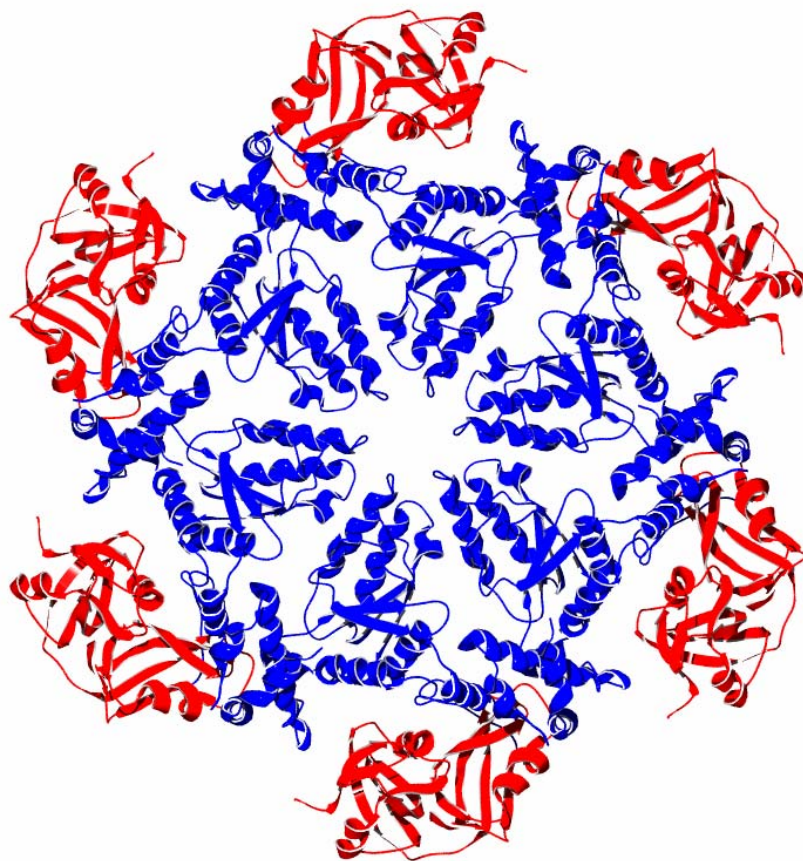
Figure 1.2 Structure of the hexameric p97 N terminal (red) and D1 (blue) domains (1E32)

Proteins of the AAA family are found in all three kingdoms of life and their utility is evident by their abundant genomic representation. Mechanistically, it is thought that these proteins exert their activity through the energy-dependent disassembly and unfolding of macromolecules. To energize such a mechanical work they use the AAA module like a chemo-mechanical converter (motor) powered by ATP hydrolysis. The conserved features of AAA domains imply a common mechanism for the operation of the 'motor'. Yet, small differences in nucleotide-sensing residues and in the domains attached to the 'motor' may determine the direction and timing of the forces that are exerted, as well as substrate specificity of these proteins. Sequence divergence in the AAA domains, followed by gene duplication in case of AAA proteins with two AAA

domains, could indeed produce an array of biological activities from a common 'ancestral' protein (Kunau et al., 1993). Further divergence in function of the AAA family comes from the wide variety of N-terminal domains. These domains act as tool heads interacting with substrates, either directly or through adaptor molecules (Dougan et al., 2002). Recently was shown that they can control also ATPase activity of the protein (Gerega et al., 2005). Extensive research on the mechanisms, which AAA proteins use for disassembly of their substrates, showed that they include processes like threading, unwinding and pulling apart of protein substrates (Maurizi and Li, 2001). Differences in the mechanism of action in proteins belonging to this family could at the end arise from differences in sequence and structure of the N-terminal domains, AAA domains or connecting regions between these two parts.

1.1.1 Phylogenetic analysis of AAA proteins

Several authors tried to analyze the phylogeny of AAA proteins based on their AAA domains over the last decade (Kunau et al., 1993; Beyer, 1997; Wolf et al., 1998; Frohlich, 2001). These analyses used different sequence subsets depending on the available data (sequenced genomes) and differed in the selection of sequences, e.g. exclusion of canonical divergent members or inclusion of degenerative, inactive and fast-evolving sequences. Even though there was some inconsistency in the selection of sequences, these analyses converged to the picture of five main clades of AAA domains in the family of AAA proteins, corresponding to proteasome subunits, metalloproteases, the loosely defined meiotic group, and D1 and D2 of the proteins with two AAA domains. Recently, a phylogenetic analysis of AAA proteins was performed with an automated approach (Frickey and Lupas, 2004). Clustering of the AAA+ superfamily, based on their ATPase domains included a well defined and compact group of AAA proteins. Filtering of these sequences was done based on their sequence conservation in Walker A and B motifs, and canonical residues in their SRH-region. Based on this approach, a phylogenetic tree was generated that comprised of six major clades. The five clades mentioned before and a new clade named BCS1, which consists of mitochondrial inner membrane proteins (Fig 1.3). Some proteins, which could not be assigned to any of major clades, form minor clades with long-branching sequences. Five of these were radiating from the root of the tree

(belphegor, Pch2p, ORF300, ORF5, and YC46) because they showed no closer relation to any major clade.
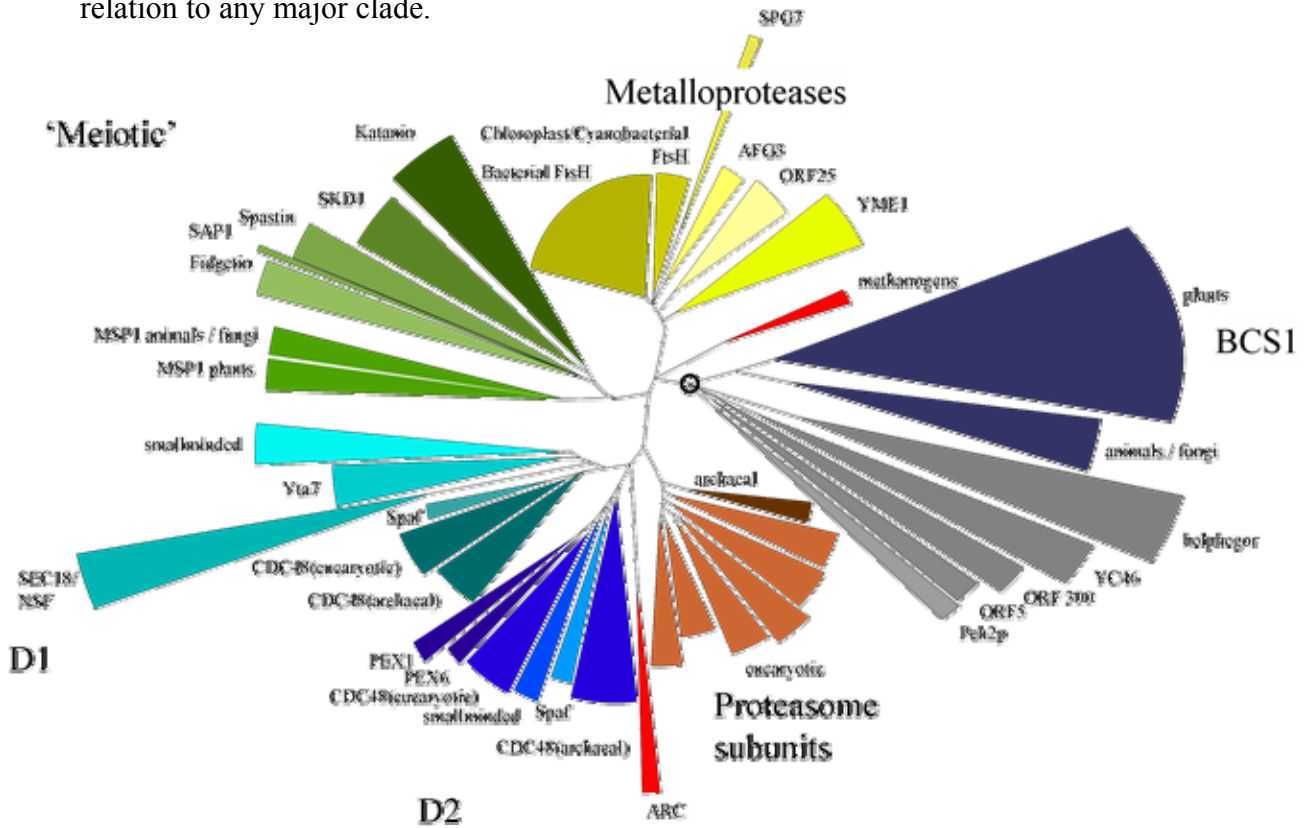
Figure 1.3 Phylogeny of AAA domains. Length and angle of each clade reflect the maximum branch length and number of sequences belonging to that clade. The presumed root of the tree is marked by a black circle and minor clades whose monophyly is unclear are colored in gray. Methanogens and ARC cluster colored in red are discussed in text (Frickey and Lupas, 2004)

Two other minor clades, even though with deep branching, were located next to major clades: ARC next to the D1 and D2 clade, and a group of methanogenic sequences next to the metalloproteases, named AMA for its occurrence in *Archaeoglobus fulgidus* and methanogenic archaea. Most of the minor clades close to the root have gapped SRH regions and are similar to the AAA+ proteins in the positions of the conserved arginine residues (arginine fingers), which may indicate an ancestral trait of these sequences. Clear clustering of the eukaryotic clades and predominance of prokaryotic sequences in deep branching clades suggests that the AAA family had already reached most of its diversity before the three domains of life separated. Such a scenario is also favorable from the significance of these proteins in the control of all major cellular processes like cell division, protein degradation, cytoskeleton, etc.

The N-terminal domains of the AAA proteins give another dimension of diversity to this group of proteins. A cluster analysis of N-terminal domains resolved these sequences into 20 groups (Appendix Fig. 1., Frickey and Lupas, 2004). This implies that the wide variety of biological functions of these proteins originates primarily from divergence in their N-terminal domains. Some of these differences were acquired after the division into plants and other eukaryotes (fungi and animals) as observed in the BCS1 clade. Others family members show high divergence in their N-terminal domains although their ATPase domains are obviously of monophyletic origin suggesting an evolutionary exchange of N-domains (like smallminded and Yta7 in the D1 clade). The most surprising findings of this cluster analysis were N-terminal homologies between the CDC48/p97 group and the deep branching AMA group as well as between ARC (AAA **A**TPase forming **R**ing shaped **C**omplexes) and proteosomal ATPases. Sequence analysis of N-terminal domains of AMA proteins using PSI-Blast (Altschul et al., 1997) and 3D-PSSM (Kelley et al., 2000) led to the assumption that these proteins adopt a structure similar to the β-clam fold found in N-terminal domains of the CDC48/p97 group. The homology between ARC and PAN proteins was proposed earlier (Wolf et al., 1998) mainly because of the presence of a coiled-coil region on the very end of the N-terminal domains and a similar genomic context (proteasome loci). Additional analysis found that significant similarity between these proteins is extended to the region between the coiled coil and the AAA domain, which is rich in β-strands and may form a β-barrel (Frickey and Lupas, 2004).

It was shown before that in AAA+ proteins the N-terminal domains either determine substrate specificity (Mogk et al., 2003) or bind to adaptor molecules which indirectly regulate their activity (Dougan et al., 2002). There are however conflicting data about the involvement of N-terminal domains in the unfoldase/chaperone activity of AAA proteins (Golbik et al., 1999; Lo et al., 2001; Beinker et al., 2002; Hinnerwisch et al., 2005). Such activity in some groups of proteins, like PAN, was not designated to any particular domain because the study was done only on full proteins (Benaroudj and Goldberg, 2000). Some new findings showed that N-terminal domains can also simply control the activity of an ATP motor module without having a defined role in substrate binding (Garega et al., 2005).

1.2 VAT protein

VAT (**V**alosine-containing-protein-like **A**TPase of *Thermoplasma acidophilum*) is an archaeal member of the CDC48/p97 group of AAA proteins and is characterized by its tripartite domain structure N-D1-D2 which is characteristic for this group (Fig.1.4). ATPase activity of VAT is Mg^{2+} dependent at an optimal temperature of approximately 70°C (Pamnani et al, 1997). Electron microscopy revealed that the members of this family, including VAT, form hexameric rings with the kidney-shaped N-terminal domain positioned at the edge of the ring structure (Fig. 1.5) (Rockel et al., 1999). These findings were further confirmed by the crystal structures of the p97 N-D1 complex (Fig. 1.2; Zhang et al, 2000) and of the complete protein (DeLaBarre and Brunger, 2005), as well as by the NMR structure of the N-terminal domain of VAT (Coles et al., 1999).

Figure 1.4 Schematic representation of the domain organization in VAT protein

In the presence of the nucleotide and Mg^{2+} the hollow cylinder with an approximate diameter of 15nm undergoes a conformational change (Rockel at al., 2002) similar to its eukaryotic homologue NSF (Fleming et al., 1998). Other eukaryotic homologues of VAT from vertebrates (p97) and yeast (CDC48) play a role in homotypic membrane fusion events (Peters et al., 1990, Latterich et al., 1995). These proteins also participate in ubiquitin–dependent protein degradation through interaction with ubiquitin fusion degradation proteins (UFDs) (Ghislain et al., 1996; Meyer et al., 2000). Furthermore, it was shown that p97 interacts with the DNA unwinding factor (DUF) which is imported into the nucleus (Yamada et al., 2000). This raised questions about an involvement of this protein in DNA replication. Nevertheless, since archaea do not contain cell organelles or ubiquitin, VAT must fulfill some different cellular function from its eukaryotic counterparts, which is up to date still unknown.
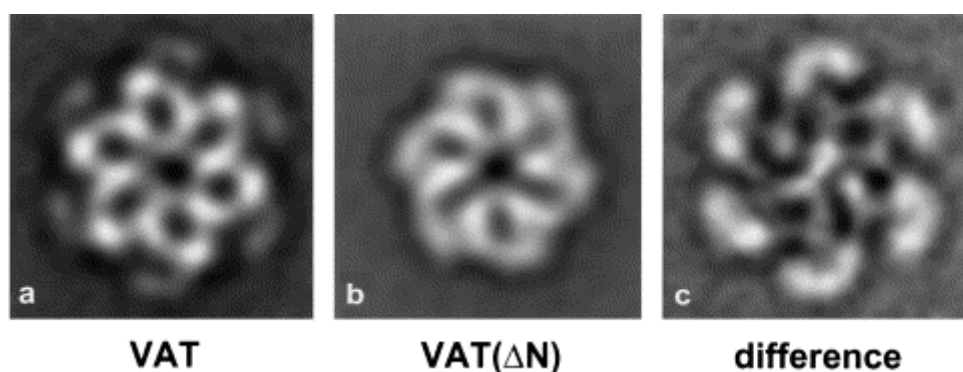
Figure 1.5 2-D average images of VAT (a) and VATΔN (b). Subtractions of the image b from image a gives clear look (c) on position and shape of the N-terminal domain (from Rockel et al., 1999)

Connection of the function of CDC48/p97 with protein degradation together with the finding that, unlike other archea, the genome of *Thermoplasma* contains no homologues of the proteasomal ATPases PAN (Proteosome Activating Nuclease) led to the hypothesis that VAT could functionally substitute PAN (Ruepp et al., 2001). This hypothesis is underlined by the fact that VAT was able to refold or unfold heterologous protein substrates *in vitro*, depending mainly on the Mg^{2+} concentration (Golbik et al., 1999). The difference between a 'low activity mode' which accelerated refolding of the non-permissive substrate penicillinase and 'high activity mode' which accelerated unfolding of the same substrate were accompanied with differences in thermal stability. The results suggested the existence of at least two different conformational states in the presence of different concentrations of Mg^{2+} ions. Moreover, the same study showed that the separately expressed N-domain of VAT possessed the ability to bind another permissive substrate, cyclophilin, and prevented its aggregation.

Recently, it was shown that VAT had Mg^{2+} dependent ATP hydrolysis and *in vitro* unfoldase activity against an ssrA-tagged GFP protein (Garega et al., 2005). Surprisingly, deletion of the N-terminal domain (VATΔN) increased ATP hydrolysis approximately 24 times, and led to an even more drastic increase (250-fold) in unfoldase activity. These data indicate the role of the kidney-shaped N-terminal domain in the regulation of activity of the full protein, but do not generally exclude the idea that this domain might have a second role in substrate or adaptor molecule

binding. Up to now the function of the VAT protein in *T. acidophilum* and the definite role of its N-terminal domain remain unclear.
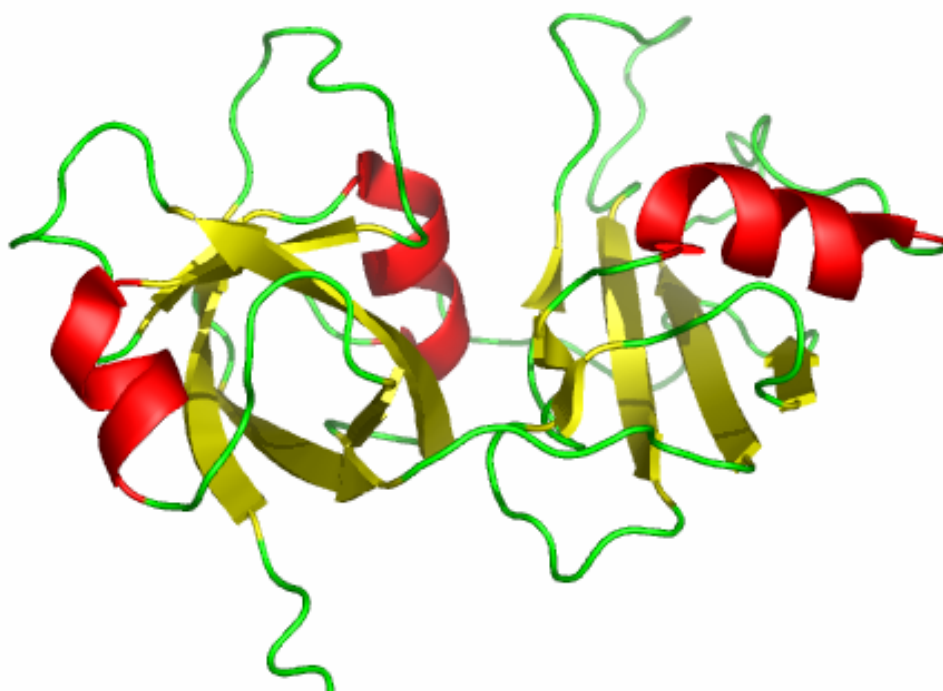
Figure 1.6 NMR structure of the N-terminal domain of VAT protein from *Thermoplasma acidophilum* (1CZ5)

1.2.1 Structure of the VAT-N domain

The structure of the N-terminal domain of VAT (VAT-N) was solved in 1999 by Coles et al. (Fig 1.6) using NMR spectroscopy. The structure contains two subdomains. The first 92 amino acid residues (named VAT-Nn) adopt a double-psi barrel fold, and the C-terminal subdomain (VAT-Nc, 93-185) folds into a so-called β-clam. VAT-N is monomeric and has the kidney-shape that was already proposed from electron microscopy data (Rockel et al, 1999). By comparison of electron microscopy data on VAT and VAT-ΔN, the position of the N-domains in the hexameric structure could be mapped (fig. 1.5). The two subdomains of VAT-N form a hydrophobic cleft that points towards the center of the hexameric ring. The orientation and hydrophobic character of the cleft perfectly match the requirements for a peptide binding pocket.

Yet, unfoldase/chaperone assays with either the VAT-N domain or with the full VAT protein (Golbik et al., 1999) never included mapping of the substrate binding site.

1.2.2 VAT-Nn

The double-psi barrel fold of the N-terminal subdomain (VAT-Nn) is one of the most complicated folds in nature. In general the fold consists of a six-stranded β barrel capped from both sides by small α-helices. Figure 1.7 shows schematically the topology of the fold so that the pseudo-twofold rotational symmetry becomes apparent. The double-psi structure is formed by two interlocked motifs, each of which comprises a loop and a strand that together resemble the Greek letter psi (Castillo et al., 1999), from which comes the name of the barrel.
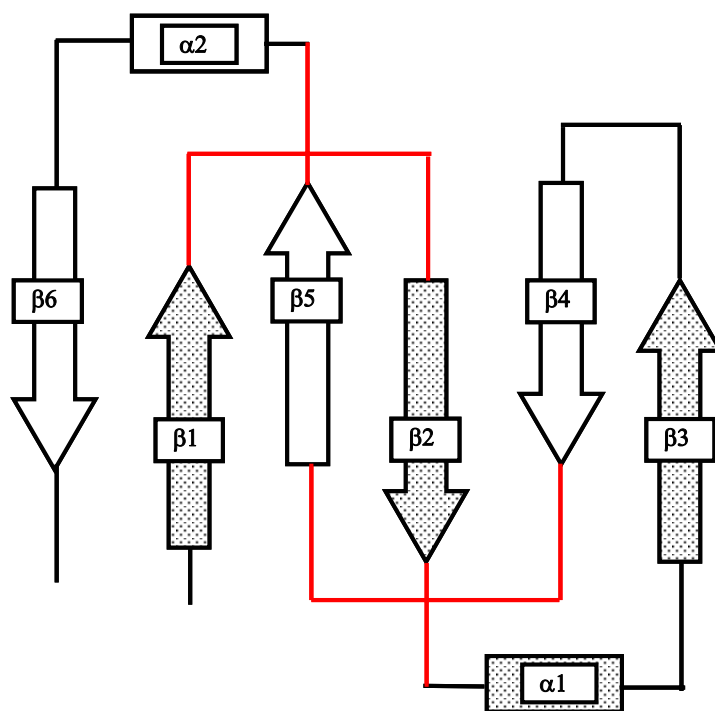
Figure 1.7 Topology of the double-psi barrel fold. The two psi-structures are shown in red color.

VAT-Nn shows a high degree of symmetry at the levels of secondary and tertiary structure, which is reflected in a sequence identity of 38% between the two halves. Two loops together with the protruding β-strands are forming a positively charged

cleft (colored blue in the surface model) that is predicted to be a binding pocket for possible substrates (Fig 1.8) (Coles et al., 1999). There is some experimental evidence from structural and functional work on the UFD1-N domain, a homologue of VAT-N, that the region between the two psi-loops is important for the binding of substrate proteins (Park et al., 2005).
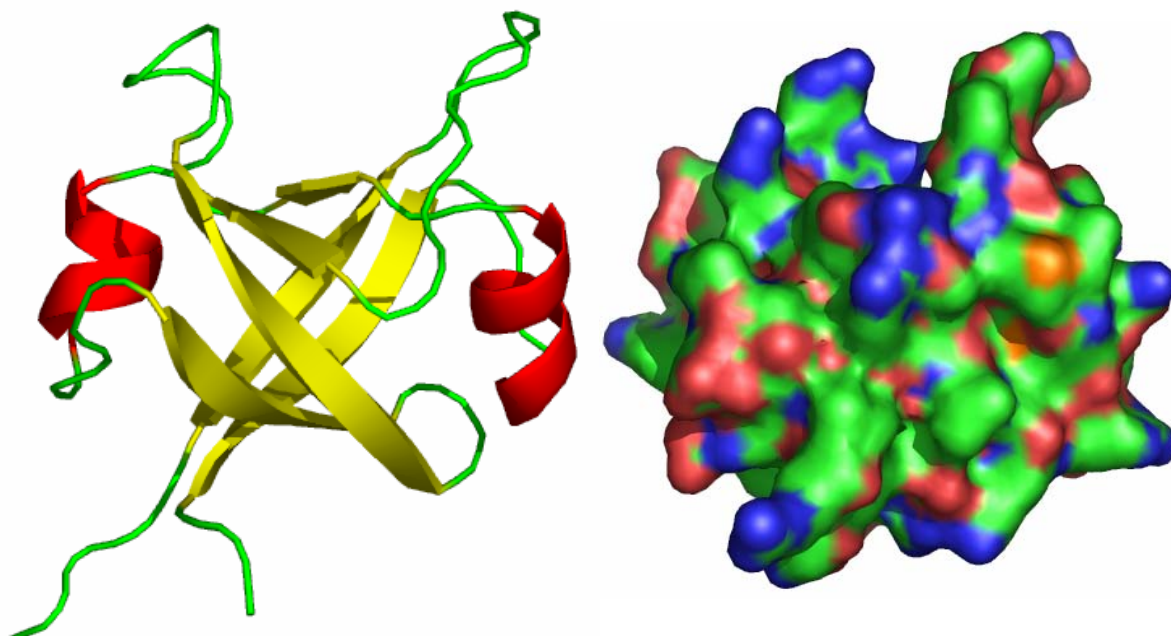
Figure 1.8 Ribbon diagram and surface model of the VAT-Nn double-psi barrel (1CZ4)

The same fold is also found in other proteins like aspartate α-decarboxylase, dimethyl sulphoxide (DMSO) reductase, endoglucanase V, in other members of the CDC48/p97 group and, with a circular permutation, in aspartic proteinases. Pseudo-twofold symmetry of the double-psi barrel fold suggests that it evolved from a homodimer by duplication (Castillo et al., 1999). While symmetry in VAT-Nn is recognizable also on the sequence level, other proteins that share the same fold keep the fold symmetry but no homology can be detected between the two halves. The most simple proteins that were found in a search with a sequence profile based on VAT-Nn repeats contained one copy of the repeat sequence and formed the barrel fold of VAT-Nn by dimerisation (Fig. 1.9) (Coles et al., 1999). There are two groups of such sequences. The first comprises archaeal transcription factors and the sequence of their repeat is entirely alignable with the VAT-Nn repeats (colored in blue Fig.

1.9). Proteins from this group have a long stretch of amino acid residues in front of this repeat which vary in length and which can raise questions of their complete structure. Another group included known and hypothetical transcription factors (colored in red Fig. 1.9) from bacteria and archaea that contain a circular permutation (Grishin, 2001) of the first β-strand to the C-terminus, forming a βαββ-element with a more simplified topology then the related VAT-Nn repeat. Based on the conservation of two beta strands that flank the α-helix and the conservation of a Gly-Asp motif (referred to as GD box), that makes an orthogonal turn in this motif, Coles et al. (1999) proposed an evolutionary path for the double-psi barrel fold.

```
                                    repeat
                      |-------------------------------------------|
                       β1        Ψ        β2   α1                β3
                      sssssss          ssss hhhhh          ssssss
VAT-Nn Ta        mesnnGIILRVAEANSTDPG-MSRVRLDESSRRLLDAE-IGDVVEIEKVr
                 KTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV                       90
UFD1 Gth        mkrleftLKTYPLSFIGKSFLENGDKIVLPQSILNYLNQNDDLNPIIFEILnldnnk
                 KCHCGVYEFTSD----DGCAYIPYWMFKNLEIN-EGSPLCFIQK                        [79]
                      sssss          ssss  hhhhh          sssss
Af1504          [121] gvDLTFTQKYDYEG----SDVAFLAEDIMNAIGVK-EGEYVSVKK
                 NGTVNLRVLPYSK---EGFIVVPTWVREKLGVK-VNDFVEVVRR                        201
                      ssss           sss  hhhhhh          ssss
PhS018            mDVLAKFHTTVHR------IGRIIIPAGTRKFYGIE-QGDFVEIKIVkyegee
                 PKEGTFTARVGE------QGSVIIPKALRDVIGIK-PGEVIEVLLLghykprn                92
                      sssss          sssss hhhhhhh         sssssss
Mj0056          [ 90] DGAIVVPKKTYHSSE-IIEIIAPMKLREQFNLK-DGDVIKILIKgdkde                 136
Mt1500          [ 75] EGAILFPLKTHHRQG-CLEFVAPVNLRKTLKLR-DGDTVSLDIDtseiqe                122
                      sssss          sss  hhhhhh         ssssssss
Mt6002757        mameDVGVPFSNRLTR----QGNIKVPADLRDALKLK-PGDLLVVEIKkvdrs                48

                      ssssss              hhhhhhh        sssssss    ssssss
Bs 1EKT/AbrBN      MKSTGIVRKVDE-----LGRVVIPIELRRTLGIA-EKDALEIYVDD—EKIILKKYkpnmt         53
Ta1217        mtdnkkiMDIARMTKRGA------SVRVTIPKKVLKKLNFK-DEDLIAFYESED-GRIYIDLLk             56
```

Figure 1.9 Alignment of the VAT-Nn sequence with related sequences. Positions in the alignment that are highly conserved or belong to the hydrophobic core are marked in red and blue, respectively; the conserved GD-box is highlighted by a red box. The secondary structure elements are indicated above the alignment; s=strand, h=helix. *Archaeoglobus fulgidus* Af1504 (gi 26409072), *Bacillus subtilis* Bs 1EKT/AbrBN(gi 113009), *Methanococcus jannaschii* Mj0056 (gi 2495770), *Pyrococcus horikoshii* Phs018 (gi 3256814), *Methanobacterium thermoautotrophicum* Mt6002757 *Thermoplasma acidophilum* Ta1217 (gi 16082225), *Guillardia theta* UFD1 Gth (gi 13812083)

This path follows the evolution of a simple 40 residue βαββ motif throughout the postulated duplication and circular permutation events first to the double-psi barrel

fold and further to the more complex fold of aspartic proteinases which comprise four similar motifs (Appendix Fig. 2). In parallel with the evolution of the fold structure, proteins function of the proteins changed from DNA binding in transcription factors to protein binding and processing in chaperones/unfoldases and aspartic proteinases. It was predicted also that members that are at the "beginning of the path" like AbrB-N and Ta1217 would fold into a barrel structure without psi-loops.

A.

B.
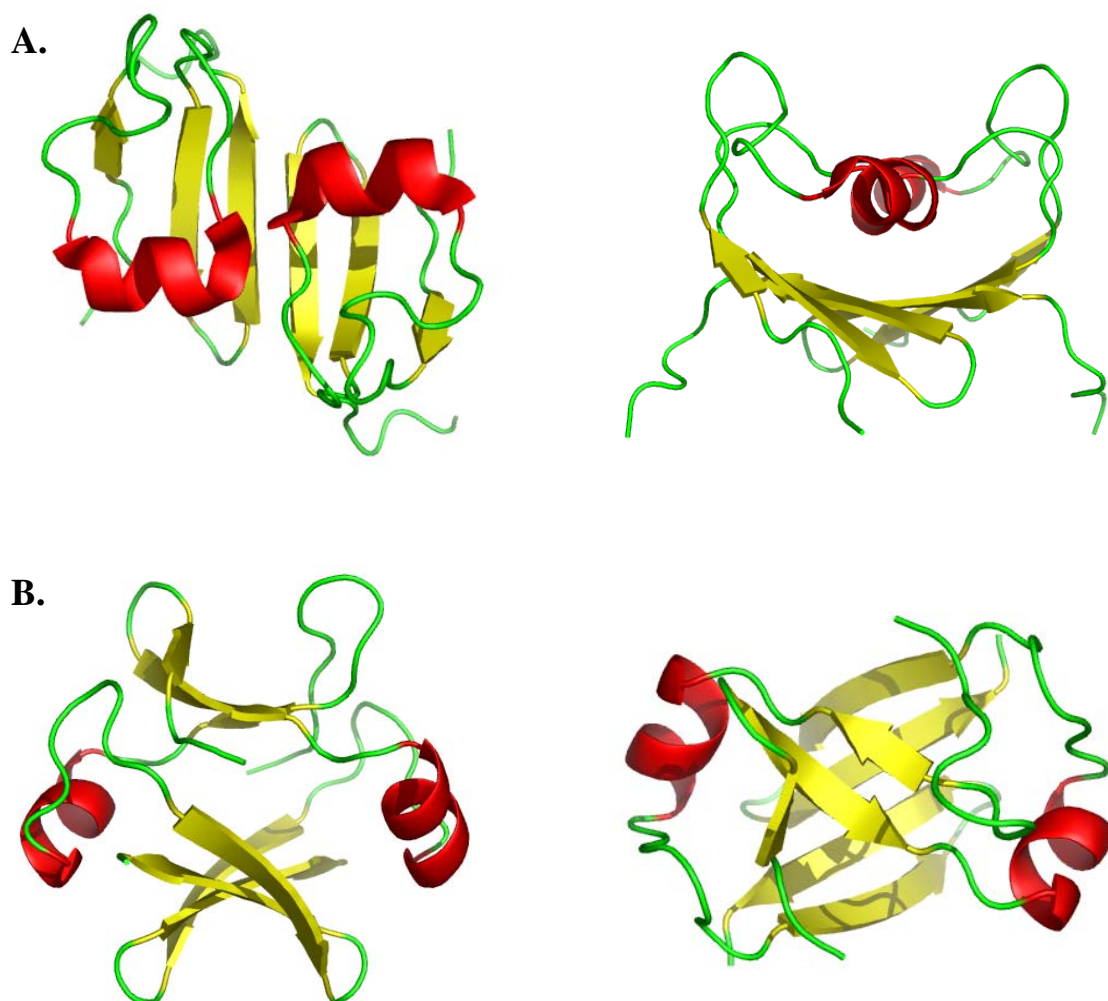
Figure 1.10 A) Ribbon diagram of the NMR structure of the AbrB transcription factor from *B. subtilis* (1EKT; top and side view); B) Crystal structure of the addiction antidote MazE a homologue of AbrB (1MVF; top and side view)

When determined by Cavanagh's group, the structure of the AbrB (Fig. 1.10 A.) transcription factor from *Bacillus subtilis* (Vaughn et al., 2000) differed greatly from the predicted barrel-like structure. The solved structure was indeed a dimer, but it was

formed side-by-side from two three-stranded beta meanders, with alpha helices placed over the beta strands and connected to them via long loops. The fold was designated new and named „looped-hinge helix". Discrepancy of the solved AbrB-N structure with the prediction (Coles et al., 1999) and moreover with the crystal structure of the homology-related MazE (1MVF) (Fig. 1.10 B.), an addiction antidote from *E. coli* (Loris et al., 2003), together with a collection of experimental evidences for the postulated evolution of the double-psi barrel fold were one starting point of this project.

1.2.3   VAT-Nc

The C-terminal part of VAT-N (VAT-Nc) adopts a so called 'β-clam' or CDC48 domain 2-like fold as defined by SCOP (Structural Classification of Proteins; http://scop.mrc-lmb.cam.ac.uk/scop/; Murzin et al., 1995). As it can be seen from Figure 1.11 its six β-strands do not form a real barrel, but rather a clam like structure that is closed on one side with an α-helix. Also noticeable is the existence of two very long loop regions between β1 and the helix and between β3 and β4 that were not resolved well in the final structure (Coles et al., 1999).
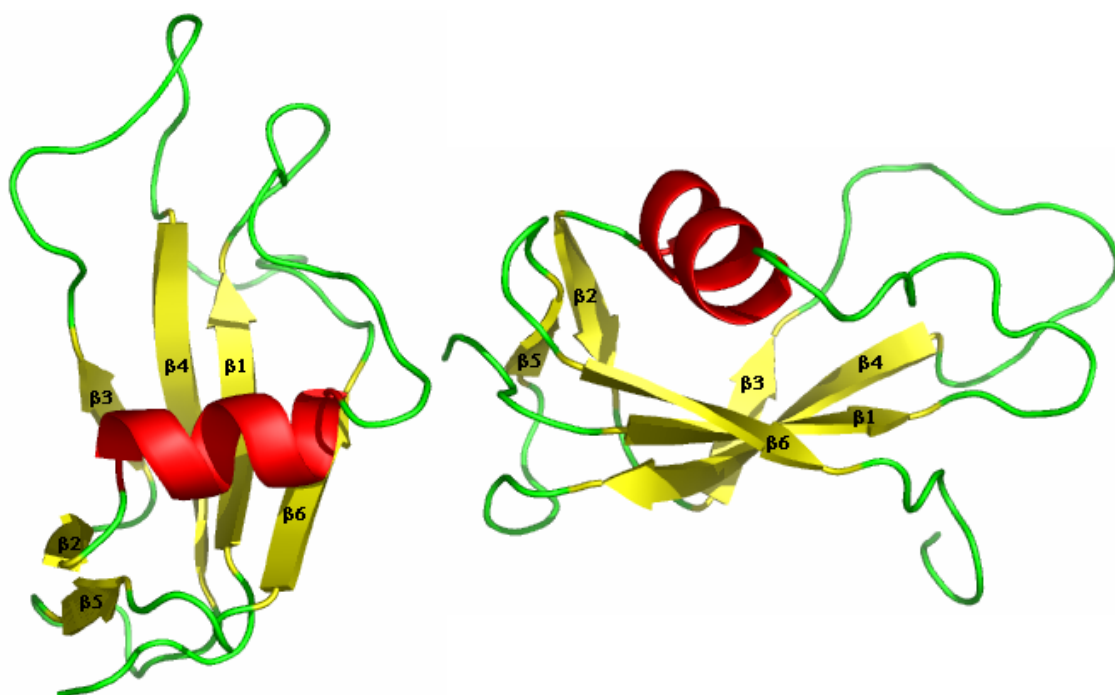
Figure 1.11 Structure of the VAT-Nc subdomain (1CZ4; top and side view). Strands forming the clam like structure are labeled β1-β6.

In the cluster analysis of proteins from the AAA family mentioned before (Frickey and Lupas, 2004), one of the minor clades, named methanogens, showed similarity in its N-terminal region with the β-clam domain of the CDC 48 group. This clade, which contains sequences from methanogenic archaea and *Archeoglobus fulgidus* (named AMA), shows a clear connection in the ATP module with the neighboring FtsH (metalloprotease) clade (Fig. 1.3). Clustering of the N-terminal domains placed sequences of the 'methanogen' group together with the CDC48 group (Appendix Fig. 1) and sequence analysis by the structure prediction server 3D-PSSM (Kelley et al., 2000) showed that these domains are likely to assume the 'β-clam' fold. AMA proteins differ from the CDC48 group in the absence of the N-terminal double-psi barrel domain and of the additional AAA module. Up to now nothing is known about this group and their domain organization. Additionally their position in the phylogenetic tree makes them interesting for experimental studies.

1.3 PAN and ARC proteins

The proteasome is a self-compartmentalizing protease, which can be found consistently in archaea and eukaryotes (Zwickl, 2002), whereas in bacteria it is found only in some members of *Actinobacteria* (De Mot et al., 1999; Zhang et al., 2004). The eukaryotic 26S proteasome degrades ubiquitin-conjugated and certain non-ubiquitinated proteins in an ATP-dependent manner (Coux et al., 1996, Voges et al., 1999). The 26S complex is composed of the 20S core proteasome surrounded by two 19S regulatory complexes, which contain six homologous ATPase subunits each. Archaea lack ubiquitin and 26S proteosome, but still their 20S proteasome resembles its eukaryotic homolog in architecture and function (Lowe et al., 1995). Proteosome Activating Nucleotidases (PANs) are archaeal homologues of the six eukaryotic ATPases in the 19S complex of the proteasome. In comparison to their eukaryotic counterparts, which form hetero-oligomeric complexes, PAN members are making homo-oligomeric complexes. Addition of ATP to a mixture of PAN complex and 20S proteasomes stimulates the formation of higher order complexes (Fig. 1.12) similar to the eukaryotic 26S proteasome, which is capable of degrading a wide variety of

protein substrates (Wilson et al., 2000). It was also shown that the PAN complex exhibits an ATP-independent chaperone activity enhanced in the presence of ATP (Benaroudj and Goldberg, 2000). Based on structure prediction PAN consists of three domains: an N-terminal coiled-coil domain followed by an β-strands rich one and an AAA domain at the C-terminus, but the structure of the protein at atomic resolution was not yet determined. Phylogenetic analysis of the AAA proteins (Frickey and Lupas, 2004) resulted in a well defined proteasomal clade that contained sequences from both archaea and eukaryotes (Fig. 1.3). Clustering of the N-terminal domains indicated a well defined group with PAN-N domains, which additionally contained N-domains of ARC proteins.
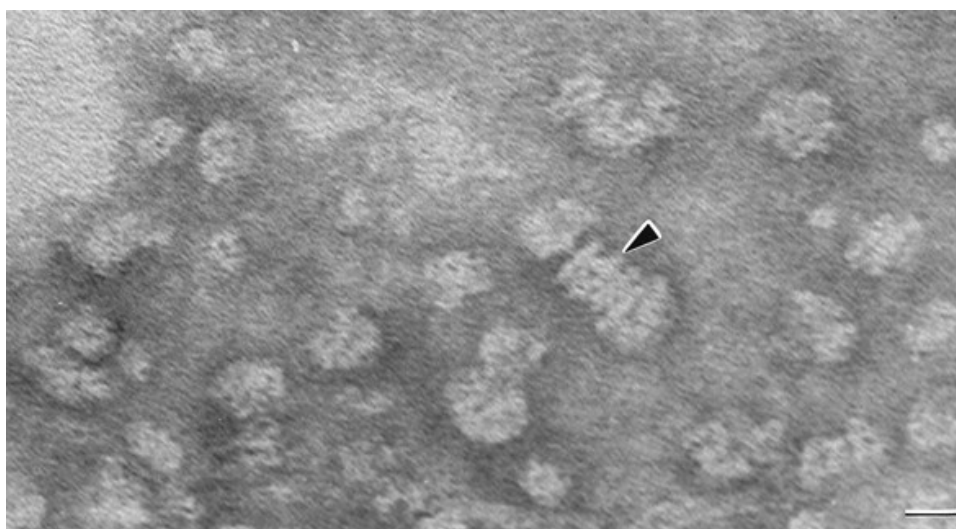
Figure 1.12 Transmission electron micrograph of reconstituted 20S proteasome and PAN proteins. The arrowhead indicates the cylindrical 20S proteasome assembly with apparent assemblies of PAN at both ends (bar represents distance of 20nm) (from Wilson et al., 2000)

When discovered in the actinomycete *Rhodococcus erythropolis* by the group of Baumeister (Wolf et al., 1998) ARC was immediately proposed to be a homolog of PAN and the 19S subunits of the proteasome. This was based on its domain composition that resembles the domain organization of PAN and its eukaryotic homologs. The fact that the gene for ARC lies in one of the proteasome operons and that no PAN-like sequences except ARC was found in the genomes of those bacteria

supports this theory. It was noticed that the same gene organization was also found in other members of Actinobacteria that contain 20S proteasomes (Wolf et al., 1998; Zhang et al., 2004). In the before mentioned cluster analysis of the N-terminal domains of AAA proteins (Frickey and Lupas, 2004), it was found that the N-domain of ARC proteins is similar with N-domains of the proteasome subunits. Similarity was not only significant in the coiled-coil segments (Wolf et al., 1998) but also in the β-strand rich interdomains, thus indicating homology. The AAA module of ARC proteins grouped in a long-branching clade between D2 and proteosome subunits clades (Fig. 1.3). When expressed alone the AAA domain of the ARC protein abolished hexamer formation. To restore the hexameric structure of the protein, the ATP domain had to be expressed covalently linked to the β-strand rich inter-domain.

1.4 Aims of this work

The general aim of this study was to perform a structural and biochemical characterization of different N-terminal domains of AAA proteins in the light of the proposed evolutionary connection between diverse members of the AAA family with homologous N-domains. In particular, their similarities were found between PAN and ARC, and between p97/CDC48/VAT and AMA, in a bioinformatic analysis of the AAA family. To elucidate these relations experimentally, N-domains of ARC and PAN as well as their subdomains were expressed and assayed for their possible chaperone activity. Structural features that might be a prerequisite for the function of these domains were tested through a set of mutations. In the course of this work we did the first characterization of *A. fulgidus* AMA AAA protein was preformed which confirmed the relation of its N-terminal β-clam domain with a homologous domain from VAT through a series of chimeric proteins.

Sequence similarity, but structural discrepancy, between VAT-Nn and the transcription factor AbrB prompted us to propose and study the origin of the double-psi barrel fold of VAT-Nn from an ancestral βαββ-motif. In search for experimental evidence for this evolutionary path we performed structural and biochemical characterization of several natural sequences closely related to either VAT-Nn or AbrB.

2. Materials

2.1 Chemicals and materials

All chemicals were ordered from Merck, Sigma Applichem, Fluka or Roth. DNA oligonucleotides were ordered from Sigma Genosys, Steinheim.

2.1.1 Escherichia coli strains

strain	relevant genotype	reference
DH10B (TOP10)	F⁻ *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*Z ΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*D139Δ (*ara, leu*)7697 *gal*U *gal*K λ⁻ *rps*L *nup*G	Grant et al., 1990; Invitrogen
C41 (DE3)	*E. coli* B F⁻ *ompT hsdSB* (r$_B$- m$_B$-) *gal dcm endA Hte* Tet^r	Miroux & Walker, 1996
BL21 gold (DE3)	*E. coli* B F⁻ *ompT hsdSB* (r$_B$- m$_B$-) *gal dcm*	Weiner et al., 1994

2.2   Buffers and solutions

2.2.1 Molecular Biology

50xTAE buffer, pH 7.6

Tris-acetate 2M

EDTA   50mM

10xTE buffer, pH 7.4

Tris Cl 100mM

EDTA   10mM

10xTBE buffer

Tris base	0.89M
Boric acid	0.89M
EDTA	20mM

Loading buffer

Xilene cyanol	0.1%	Bromphenol blue	0.1%
SDS	0.5%	EDTA,pH 8.0	0.1M
Glycerol	50%		

Bacterial culture media:

LB agar

1.5% Agar in LB broth

LB broth (1l), pH 7.0

LB-Kan (Amp, Chl) agar

Bactotriptone	10g
Yeast extract	5g
NaCl	10g

50μg Kan in ml of LB agar

100μg Amp in ml of LB agar

37μg Chl in ml of LB agar

Antibiotics stock solution

50mg/ml of Kanamycin in dH_2O

10xM9 minimal medium salts (per liter)

100mg/ml of Ampicilin in dH_2O

68 g	Na_2HPO_4
30 g	KH_2PO_4
5 g	NaCl

37mg/ml of Chloramphenicol in EtOH

100xVitamin solution (per liter)

100 mg	Choline chloride	100 mg	Nicotinamide
100 mg	D-Calcium pantothenate	100 mg	Pyridoxal hydrochloride
100 mg	Folic Acid	10 mg	Riboflavin
200 mg	i-Inositol	100 mg	Thiamine hydrochloride

2.2.2 Protein Biochemistry

SDS PAG-Electrophoresis

Running gel buffer	**Stacking gel buffer**	**Electrophoresis buffer**
1.5 M TRIS/HCl, pH 8.8	0.5 M TRIS/HCl, pH=6.8	25 mM TRIS/HCl, pH=8.3
0.4 % SDS	0.4% SDS	192 mM glycin

Sample buffer

130 mM	TRIS/HCl, pH 6.8
10%	SDS
10%	β-mercaptoethanol
20%	glycerol
0.06%	bromphenolblue
2%	LiDS

Coomassie staining solution

0.2 % Coomassie Brilliant-Blue G 250

10 % acetic acid & 50 % methanol

Destaining solution

10 % acetic acid

30 % EtOH

His-tagg protein purification

Lysis (loading) buffer (A)

30 mM   TRIS/HCl, pH 7.9

300 mM NaCl

Washing buffer (B)

30 mM   TRIS/HCl, pH 7.9

300 mM NaCl

50 mM   Imidazole

Elution buffer (C)

30 mM   TRIS/HCl, pH 7.9

300 mM NaCl

1 M   Imidazole

Loading buffer* (A1)

50 mM   TRIS/HCl, pH 7.9

30 mM   Na-phosphate, pH 7.9

8M   Urea

Washing buffer* (B1)

50 mM   TRIS/HCl

30 mM   Na-phosphate, pH 6.2

8M   Urea

Elution buffer* (C1)

50 mM   TRIS/HCl

30 mM   Na-phosphate, pH 4.7

8 M   Urea

*Under denaturing conditions

1xPBS buffer (per liter)

8 g   NaCl

1.15 g Na_2HPO_4

0.2 g   KCl

0.2 g KH_2PO_4

Chaperone assay buffer

30mM   HEPES, pH 7.2

+ 5mM $MgCl_2$

+120mM KCl

GST elution buffer (D)

50 mM TRIS/HCl pH 8.0

10 mM glutathione

30xthrombin cleavage buffer (TCB)

1.5 M   TRIS/HCl pH 8.0

3 M   NaCl

75 mM $CaCl_2$

Phenylmethansulfonylchloride(PMSF)   2.2 mg/ml   (stock)

BCA assay Solution A (Pierce)

Contains BCA, sodium carbonate, sodium tartrate and sodium bicarbonate in 0.1 M NaOH (pH 11.25)

BCA assay Solution B

4% $CuSO_4$ x $5H_2O$

Bradford reagent

100 mg   Coomassie Brilliant Blue G-250

50 ml   Ethanol (95%)

100 ml   phosphoric acid (w/v) (85%)

water up to 1 l

Western blot

Towbin buffer

  25 mM   TRIS-HCl, pH 7.5

 192 mM   Glycin

    29 %   Methanol

0.0375 %   SDS

TBS-buffer

 50 mM   TRIS-HCl

150 mM   NaCl

M-TBS

5 %   milk powder in TBS buffer

TBS-T

 0.1 %   Tween 20 in TBS

Malachite Green Solution

0.3 g Malachite Green Oxalate

2   g Sodium Molybdate

0.5 g Triton X-100

Bring to 1 liter total volume with 0.7 M HCl. Store in dark bottle at 4°C

2.3 Plasmids

Xho I(158)
Not I(166)
Eag I(166)
Hind III(173)
Sal I(179)
Sac I(190)
EcoR I(192)
BamH I(198)
EcoR V(206)
Nco I(212)
Kpn I(238)
Bgl II(241)
Nsp V(268)
Nde I(346)
Xba I(384)
SgrA I(495)
Sph I(651)
Bpu1102 I(80)
Dra III(5180)
f1 origin (4956-5411)
Pvu I(4479)
Sgf I(4479)
Sma I(4353)
kan (4048-4860)
Nru I(4136)
pET30b
Mlu I(1176)
Bcl I(1190)
lacI (826-1905)
BstE II(1357)
Apa I(1387)
BssH II(1587)
Hpa I(1682)
Eco57 I(3825)
AlwN I(3693)
ori (3339)
BssS I(3450)
BspLU11 I(3277)
Sap I(3161)
Bst1107 I(3048)
Tth111 I(3022)
PshA I(2021)
Bgl I(2240)
Fsp I(2258)
Psp5 II(2283)

Source: Novagen

Xho I(158)
Not I(166)
Eag I(166)
Hind III(173)
Sal I(179)
Sac I(190)
EcoR I(192)
BamH I(198)
Nhe I(231)
Nde I(238)
Nco I(296)
Xba I(335)
Bgl II(401)
SgrA I(442)
Sph I(598)
Bpu1102 I(80)
Dra III(5127)
f1 origin (4903-5358)
Pvu I(4426)
Sgf I(4426)
Sma I(4300)
Kan (3995-4807)
Cla I(4117)
Nru I(4083)
pET28b
Mlu I(1123)
Bcl I(1137)
lacI (773-1852)
BstE II(1304)
Apa I(1334)
BssH II(1534)
EcoR V(1573)
Hpa I(1629)
Eco57 I(3772)
AlwN I(3640)
ori (3286)
BssS I(3397)
BspLU11 I(3224)
Sap I(3108)
Bst1107 I(2995)
Tth111 I(2969)
PshA I(1968)
Bgl I(2187)
Fsp I(2205)
Psp5 II(2230)

pGEX-4T-1 (27-4580-01)

Thrombin

| Leu | Val | Pro | Arg | Gly | Ser | Pro | Glu | Phe | Pro | Gly | Arg | Leu | Glu | Arg | Pro | His | Arg | Asp |
| CTG | GTT | CCG | CGT | GGA | TCC | CCG | GAA | TTC | CCG | GGT | CGA | CTC | GAG | CGG | CCG | CAT | CGT GAC | TGA |

BamH I   EcoR I   Sma I   Sal I   Xho I   Not I   Stop codons

glutathione S-transferase

Bal I
Tth111 I
Aat II

Ptac
BspM I

pSj10ΔBam7Stop7

Amp^r

Pst I

pGEX
~4900 bp

Nar I
EcoR V
BssH II
Apa I
BstE II
Mlu I

lac I q

p4.5

AlwN I

pBR322
ori

Source: Amersham

NdeI   EcoRI   SmaI   BamHI   SalI   HindIII   ClaI

MCS

PT7

pT7-7 vector

Amp r

ColE1 ori
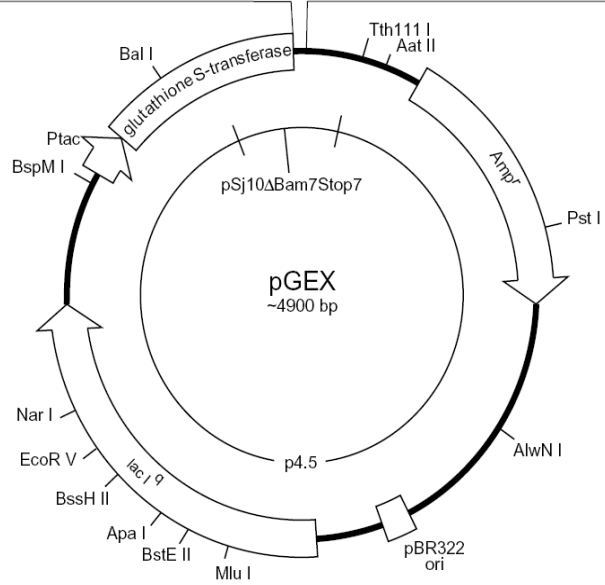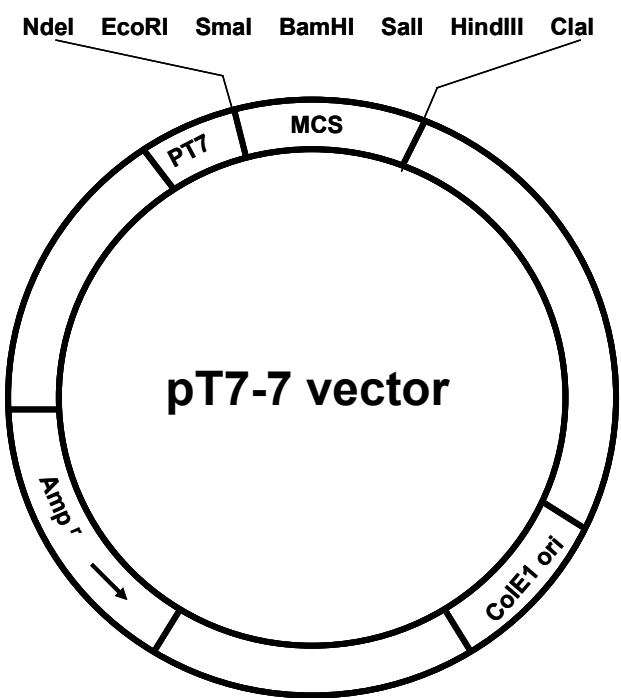
Source: Invitrogen
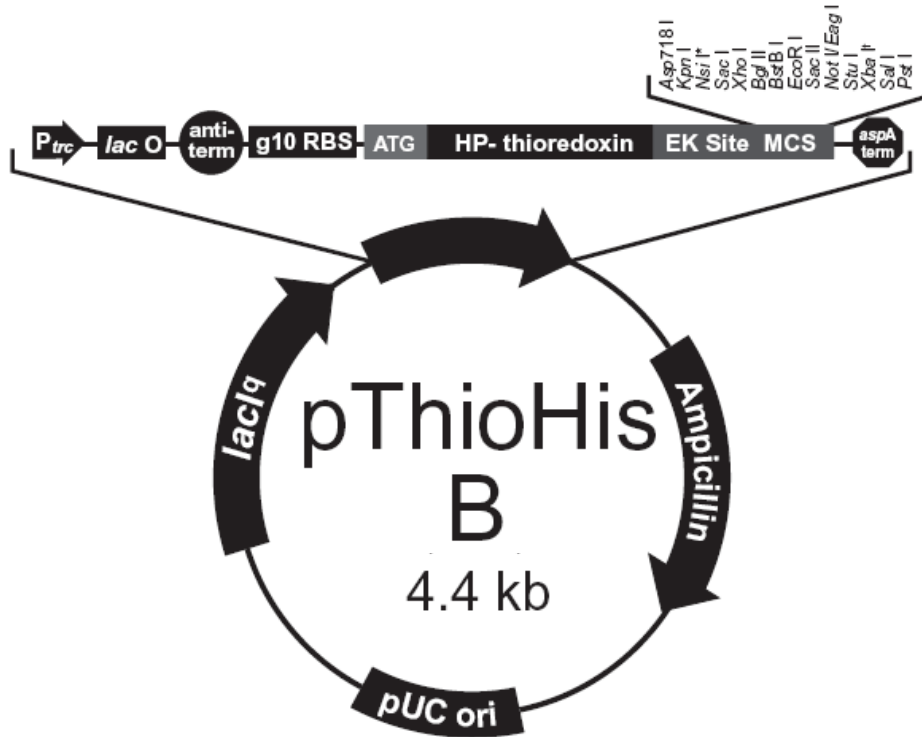
2.4 Oligonucleotides

Name	Sense	Sequence (5'→ 3')	Comment
Sequencing primers for plasmids			
T7 promotor	s	TAA TAC GAC TCA CTA TAG GG	pET30, pET28
T7 terminator	as	GCT AGT TAT TGC TCA GCG G	pET30, pET28
pGEX prom.	s	CCA GCA AGT ATA TAG CAT GG	pGEX-4T-1
pGEX termin.	as	CCG GGA GCT AGA GCA TGT GTC GG	pGEX-4T-1
VAT-Nn project primers			
Fvatnn1	s	**TCT AGA** TAA **GAA GGA G**TA **TAG GAT CC**A TGG AAT CAA ACA ACG G	XbaI, Rbs, BamHI
Rvatnn1	as	**CTC GAG CAT ATG GAT CC**A T**CT CCT TC**T TGT GAT TAA CGA ACT TTT TCG AT	XhoI, NdeI, BamHI, Rbs
Fvatnn1db	s	GC**T CTA GA**A A**GA AGG AG**A TAT A**GG ATC C**AT GGA ATC AAA CAA CGG	XbaI, Rbs, BamHI
Rvatnn1db	as	GGG AAT TC**C ATA TG**A CGA ACT TTT TCG AT	NdeI
Rvatnn1db	as	GGG AAT TC**C ATA TG**A CGA ACT TTT TCG AT	NdeI

Name	Sense	Sequence (5'→ 3')	Comment
Fvatnn1nw	s	GGA ATT C**CA TAT G**AA CGG TAT CAT CCT CCG TGT TGC A	NdeI
Rvatnn1nw	as	CCC **AAG CTT** AAC GAA CTT TTT CGA T	HindIII
R-per1va-	as	CCG **CTC GAG** TTA TGT CAT GTT TGG TTT GTA TTT TTT TAG TAT TAT TTT TTC ACG AAC TTT TTC G	XhoI
F-vatnn2	s	CGC **CAT ATG** GTT CGT AAA ACT GTT GGT CGT G	NdeI
F-gstvatnn2	s	CG**G GAT CC**G TTC GTA AAA CTG TTG GTC	BamHI
R-gstvatnn2	as	GG**C TCG AG**T TAA ACT TTA CGA ACT TTT AC	XhoI
Rvatnn2his	as	CGC CCC **AAG CTT** AAT GAT GAT GAT GAT GAT GGC CGC CAA CTT TAC GAA C	HindIII
Rvatnn2	as	CCC **AAG CTT** AAA CTT TAC GAA CTT TTA C	HindIII
VnnM F 1	s	ATG GAA TCA AAC AAC GGT ATC ATC CTC CGT GTT GCA GAA GCA AAC TC AA CTG ATC CG	
VnnM R 1	as	AAG CAG ACG TCG CGA TGA TTC ATC AAC ACG GAT ACG TGA CAT ACC CGG ATC AGT TGA G	
VnnM F 2	s	GTT GAT GAA TCA TCG CGA CGT CTG CTT GAT GCA GAA ATC GGT GAT GTA GTT GAA ATC G	
VnnM R 2	as	ACG TGC ACG ATA AAC ACG ACC AAC AGT TTT ACG AAC TTT TTC GAT TTC AAC TAC ATC	
VnnM F 3	s	GGT CGT GTT TAT CGT GCA CGT CCT GAA GAC GAA AAT AAA GGT ATT CTC CGT GTT GAC	

Name	Sense	Sequence (5'→ 3')	Comment
VnnM R 3	as	CCG CAA TTA TTA CGC ATA ACT GAG TCA ACA CGG AGA ATA CC	
VnnM F 4	s	TCA GTT ATG CGT AAT AAT TGC GGT GCA TCA ATC GGT GAC AAA GTA AAA GTT CG	
VnnM R4	as	GCC C**AA GCT T**AA ACT TTA CGA ACT TTT ACT TTG TCA CC	HindIII
VnnMBeta R1	as	GAT TCC GGG ATG ATG ATA CGA CCC ATA CCC GGA TCA GTT GAG TTT GCT TCT GC	
VnnMBeta F2	s	GGT ATG GGT CGT ATC ATC ATC CCG GAA TCA TCG CGA CGT CTG CTT GAT GCA GAA ATC G	
VatwlF1	s	GGA ATT C**CA TAT G**GA ATC AAA CAA CGG TAT CAT CCT CCG TGT TGC AGA AGC AAA CAT GT	NdeI
VatwlF2	s	CCG TGT TGC AGA AGC AAA CAT GTC ACG TGT TCG TCT CGA TG	
VatwlF3	s	GTT TAT CGT GCA CGT AAA GGT ATT GTT CGT ATT GAC TCA G	
VatwlgstF	s	CGC **GGA TCC** GAA TCA AAC AAC GGT ATC ATC CTC CG	BamHI
VatwlgstR	as	GCC GCT CGA GTT AAA CTT TAC GAA CTT TTA CTT TGT CAC C	XhoI

Name	Sense	Sequence (5'→ 3')	Comment
2ndwlF1	s	GGA ATT C**CA TAT G**AA AAC TGT TGG TCG TGT TTA TCG TGC ACG TAA AGG TAT TGT TCG	NdeI
2ndwlF2	s	CGT GCA CGT AAA GGT ATT GTT CGT ATT GAC	
2ndwlR	as	GCC C**AA GCT T**AA ACT TTA CGA ACT TTT ACT TTG TCA CCG	HindIII
2ndwlRdb	as	CTG TTT TGC GTG GTT CTT CGC CCT CAA CTT TAC GAA CTT TTA CTT TGT CAC C	
Cper2F	s	GGA ATT CCA TAT GGC ACG TCC TGA AGA CGA AAA TAA AGG	
Name	**Sense**	**Sequence (5'→ 3')**	**Comment**
Cper2R1	as	AAC TTT ACG AAC TTT TAC TTT GTC ACC G	
Cper2R2	as	GCC C**AA GCT T**AA CGA TAA ACA CGA CCA ACA GTT TTA CGA ACT TTA CGA AC	HindIII
CperCR	as	GCC C**AA GCT T**AA CGA TAA ACA CGA CCA ACG CAT TTA CGA ACT TTA CGA AC	HindIII
VnncbetaTaR	as	CC**A AGC TT**A TTT GAG GAG GTC GAT ATA GAT ACG ACC GTC TTC AAC TTT ACG AAC TTT TAC	
DbNncwl F1	s	CAA TCG GTG ACA AAG TAA AAG TTC GTA AAG TTG AGG GCG AAG AAC CAC GCA AAA CAG TAG	
DbNncwl R1	as	ACA CTA TCT ATT CTT ACT ATT CCT TTT CTT GCT CTA TAT ACT CTT CCT ACT GTT TTG CG	
DbNncwl F2	s	GAA AAG GAA TAG TAA GAA TAG ATA GTG TAA TGA GAA ATA ATT GTG GAG CAA GTA TAG G	
DbNncwl R2	as	GCC C**AA GCT T**AT ACT TTT CTT ACT TTT ACT TTA TCT CCT ATA CTT GCT C	HindIII

VAT-Nn related project primers			
Wid F1	s	GGA ATT CC**C ATA TG**G TTC GTA AAA CAG TTG GTC GTG TTT ATC GTG CAC GTC CTG AAG ACG	NdeI
Wid R1	as	CCT AAT TCA TCA ACT TTA CGT ACA ATA CCA GTA GAT TTG CCG TTA ACT TTA CGA ACT TTT ACT TTG TCA CCG	
Wid F2	s	CGG TGA CAA AGT AAA AGT TCG TAA AGT TAA CGG CAA ATC TAC TGG TAT TGT ACG TAA AGT TGA TGA ATT AGG	
Name	**Sense**	**Sequence (5'→ 3')**	**Comment**
Ta1217-F1	s	TGG C**CA TAT G**CA TCA TCA TCA TCA TCA TGG CGG TAC TGA CAA CAA A	NdeI
Ta1217-R1	as	GGT CAT GCG GGC GAT ATC CAT AAT TTT TTT GTT GTC AGT AC	
Ta1217-F2	s	CCG CAT GAC CAA ACG CGG TGC TAG CGT CCG TGT AAC CAT TCC G	
Ta1217-R2	as	TTC ATC TTT GAA GTT CAG TTT TTT CAG TAC TTT TTT CGG AAT GGT TAC	
Ta1217-F3	s	ACT TCA AAG ATG AAG ATC TGA TTG CAT TTT ATG AAA GCG AAG ATG GCC	
Ta1217-R3	as	GCC **AAG CTT** ATT TCA GCA GGT CGA TGT AAA TGC GGC CAT CTT CG	HindIII
Mt1500gst-F1	s	**GGA TCC** GAA GGC GCC ATC CTG TTC	BamHI
Mt1500-F1	s	CAT CG**C ATA TG**C ATC ATC ATC ATC ATC ATG GTG GTG AAG GCG CC	NdeI
Mt1500-R1	as	CGG TGA TGG GTT TTC AGC GGG AAC AGG ATG GCG CCT TCA	
Mt1500-F2	s	ATC CTG TTC CCG CTG AAA ACC CAT CAC CGC CAG GGC TGT CTG GAA TTT GTG GCG CCG GTG A	
Mt1500-R2	as	CGC GCA GTT TCA GGG TTT TAC GCA GGT TCA CCG GCG C	
Mt1500-F3	s	CCT GCG TAA AAC CCT GAA ACT GCG CGA TGG CGA TAC CGT GAG CCT GGA	

Name	Sense	Sequence (5'→ 3')	Comment
		TAT CGA CAC C	
Mt1500-R3	as	TGC **AAG CTT** ATT CCT GAA TTT CGC TGG TGT CGA TAT CC	HindIII
Mt6002757gst-F1	s	**GGA TCC** GCG ATG GAG GAT GTT GGC	BamHI
Mt6002757-F1	s	ATG T**CA TAT G**CA TCA TCA TCA TCA TCA TGG CGG CGC GAT GGA GG	NdeI
Mt6002757-R1	as	GGC GGT TGC TGA ACG GCA CGC CAA CAT CCT CCA TCG C	
Name	**Sense**	**Sequence (5'→ 3')**	**Comment**
Mt6002757-F2	s	GTT GGC GTG CCG TTC AGC AAC CGC CTG ACC CGC CAG GGC AAC ATT AAA GTG CCG	
Mt6002757-R2	as	CAG TTT CAG CGC ATC ACG CAG GTC CGC CGG CAC TTT AAT G	
Mt6002757-R3	as	TGC **AAG CTT** AGC TGC GGT CAA CTT TTT TGA TCT CCA CC	HindIII
Mt6002757-F3	s	CGG ACC TGC GTG ATG CGC TGA AAC TGA AAC CGG GTG ACC TGC TGG TGG TGG AGA TCA AA	
MT60errF	s	ATG GCG ATG GAG GAT GTT GGC GTG CCG TTC AGC AAC CGC CTG ACC CGC CAG GGC	
Mt60gstR	as	GGC **GAG CTC** TTA GCT GCG GTC AAC TTT TTT GAT CTC CAC C	XhoI
Mt60newF1	s	GGA ATT C**CA TAT G**GC GAT GGA GGA TGT TGG CGT GCC G	NdeI
Mt60dbR1	as	GGA ATT C**CA TAT G**GC CGC TGC GGT CAA C	NdeI
MT60dbgstF	s	CGC **GGA TCC** GCG ATG GAG GAT GTT GGC G	BamHI
MT60dbgstR	as	CCG **CTC GAG** TTA GCT GCG GTC AAC TTT TTT GAT CTC CAC C	XhoI
Mt60dbF	s	**TCT AGA** AA**G AAG GAG AG**T CCG AGC TT A TGG CGA TGG AGG ATG	XbaI, Rbs

PhF1	s	CGC **CAT ATG** CAG AAC CAG CAG AAA ACC GTT GAA CCG CTG GCT AAA TTC CAT GCT	NdeI
PhR1	as	AAC CGG AAC AAC CAG CTG ACC TTT GAT GTT AAC GCT AGC ATG GAA TTT	
PhF2	s	GGT CAG CTG GTT GTT CCG GTT AAA GAC CGT GAA GTT TTC GGT CTG AAA CGT GGT GAC ATC CTG GAA ATC ATC GTT	
PhR2	as	GAT ATG GAT TTT ACC GTT GAT AAC GTC GAA GCT ACG AAC GAT GAT TTC C	
Name	**Sense**	**Sequence (5'→ 3')**	**Comment**
PhF3	s	GTT ATC AAC GGT AAA ATC CAT ATC AAA AAA CGT GCT TAC ATC CTG GTT CGT CTG AGC AGC AAA GGT CTG ATC ACC ATC	
PhR3	as	CGG GCT GAT ACC CAG TTC ACG ACG AAC TTC TTC CGG GAT GGT GAT CAG	
PhF4	s	CGT CGT GAA CTG GGT ATC AGC CCG GGT GAC ACC GTT GAA GTT CTG CTG GTT GGT TTC CAT AAA TTC GAC GAA CTG	
PhR4	as	GAT CAG TTT AGC GAT CTG TTT ACC TTT TTC GGT AAC CAG TTC GTC GAA	
PhF5	s	GGT AAA CAG ATC GCT AAA CTG ATC CAG GCT AAC ACC CAC ATG CGT CTG ATC ACC AGC GAA GAA GAA AAA ACC ATC	
PhR5	as	CCC **AAG CTT** AAA CGT AGT AGG TAC GGC TTT TTT CGA TGA TGG TTT TTT C	HindIII
MjvatF	s	GGA ATT C**CA TAT G**GT GAA ATT GAT GAT TAT TGA GGG AGA AGT AGT TTC AGG	NdeI
MjvatR	as	GCC C**AA GCT T**AT TCA TCT TTA TCT CCC TTA ATT AGT ATT TTT ATA AC	HindIII
AbrB and related project primers			
AB F	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TTT TAT GAA ATC TAC TGG	NdeI

Name	Sense	Sequence (5'→ 3')	Comment
		TAT TGT ACG	
AB R	as	GCC C**AA GCT T**AG TTT GGT TTA TAT TTT TTA AGG ATG ATT TTT TCA TCA TC	HindIII
SpF	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TAA AGC AAC CGG TAT CGT ACG TCG TAT TGA TGA CTT AGG	NdeI
SpR	as	CCG **CTC GAG** TTA CTG TTC CAT TTG ACG AGC CAA AAA TCC AGC TGC TGT TTC AAC GGC	XhoI
Name	**Sense**	**Sequence (5'→ 3')**	**Comment**
SpcF	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TGG AGA CTT TGC AAA GGA GTA TGC AGA CGC GCT TTA CGA CAG CC	NdeI
forSpVOp	s	CGT AAA TTT TGT TAC TCT CTG GTG TAT ATT ACA TTT GAT GTG ACG GAT ACT AAT TTC AAG CGA GGC GGA AGG TAC ATA AAG TAA CTG CTT AGG TTC TTT	
revSpVOp	as	AAT ACG ACG TAC GAT ACC GGT TGC TTT CAT CTC TGG TGC CTC TCT TTC ATT TGA TGG TAT ATA CAT GTG GGA AAG ACC TAA AGC AGT TAC TTT ATG TAC	
Bofc4For	s	GGT GCT CAT GAT TAT CGC CTG CTG CTA CGG CGC GGC TTT AGC AGG G	
Bofc2For	s	GGC GTG ATT TGG ATC TTT GCG GCA TTG TAT TTT TTC AGT GCC TTT CTG ACG	
BofcRev	as	GCC AAG TGC GCG GGA AGG AGC TCC TAT TAA AAA TAC TGC CGC	
Per2F1	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TGC GGG CCA AGA GCA GCA GGT TCG CGT G	NdeI
Per2R1	as	CAT CAC GGT CAA TGC CGG CAC GTT TCA GCG CAG AAA CCG GAA TAG ACA CGC GAA CCT GC	
Per2F2	s	CGG CAT TGA CCG TGA TGA TCA GCT GCA GCT GAA AGT TCG CAA CGA CGC	

Name	Sense	Sequence (5'→ 3')	Comment
		CTT AGT TCT GG	
Per2R2	as	**AAG CTT** AGT ACT GTT GCC ATT GCG GCT GGC GCG GTT TAA TTT TCT CCA GAA CTA AGG CGT CGT TGC G	HindIII
Per1F1	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TTC TAA AAC CCC GTA CCA GCT GAC CCT	NdeI
Name	**Sense**	**Sequence (5'→ 3')**	**Comment**
Per2Rshort	as	CCC **AAG CTT** ACT GGC GCG GTT TAA TTT TCT CCA GAA CTA AGG	HindIII
Per1R1	as	GCT GGT TAT GTT TCA GGT GCA GTT CAG ACA CAA TTT CAT CCG GCA GGG TCA GCT GGT ACG	
Per1F2	s	GCA CCT GAA ACA TAA CCA GCC GCT GAA CCT GAC CCT GCG TAA CGG CCA GCT GAT CAT TC	
Per1R2	as	GCC C**AA GCT T**AG GTC TGG CGC TCG TTC GGA TTG TTG GTC TGT TGA ATG ATC AGC TGG C	HindIII
F1op	s	CTT CAT CAA TCG TTC CTC CTC ATA ATT GTG GTC TCC CTG TTT AGG GTA CCT GAA ATT AAC AGT CAG GAC AAT CAA TAT ATC ATA TCT TTT GC	
R1op	as	CAT AAT TTA TCA GAG CGG AGG AAA GAT ATT GCT CCC ACA CAT ACT GAA TTA TTA TCA TCG TAA TCT GTT TTT GCA AAA GAT ATG ATA TAT TGA TTG TCC	
F2op	s	CAG TAT GTG TGG GAG CAA TAT CTT TCC TCC GCT CTG ATA AAT TAT GTA AAG ATA CAA ATT GAT TTA TAT AAA GCG CTT AAA AAT TAG TAT AAG	
R2op	as	GTACAAGTATTATATCATCTTGGATGTTCCGTGAAAAAAATACTGTATAGTAAATAACTTATACTAATTTTTAA GCGCTTTATATAAA	
F3op	s	CAC GGA ACA TCC AAG ATG ATA TAA TAC TTG TAC TAA ATA AAT ACA AAT CGC ATT ATG CCT TCT GTA TTT ATA AAC ATC AAT TTT CAT CTC GAG	

R3op	as	CGA TAT TTC TTA TAG ATG AAA TGA ATT GAA CGG CGG GCA AAT ATT TGC AAT ATT AAA CGA CAG GTA GTA TCA ATT TTT TCT CGA GAT GAA AAT TGA TG	

Name	Sense	Sequence (5'→ 3')	Comment
VAT-Nc and related project primers			
V-NCF	s	GGG AAT TC**C ATA TG**C GTA CCG AAA TCG CTA AAA AGG TTA CTC TGG CA	NdeI
V-NCR	as	GCC C**AA GCT T**AA TGG TGA TGG TGA TGG TGA CCA CCT TCC TCG AGA AC	HindIII
RV-NCwH	as	GCC C**AA GCT T**AT TCC TCG AGA ACT TCT GAA GCC GG	HindIII
AfclF	s	GGA ATT C**CA TAT G**GC AAA GAG GGA AAC GGC TGA GTT GAG ATA CCT GAT AG	NdeI
AfclhisF	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TGC AAA GAG GGA AAC GGC TGA G	NdeI
AfclR	as	GCC C**AA GCT T**AG TCT CTG ACA ATT TCG GTC TCG ATA ACT CTG GGG TCG CTT TCC	HindIII
MjclF	s	GGA ATT C**CA TAT G**AG TAA AAT TGG ATT TAA TCC AAT AAA AAT AAA ATC TTT TTC	NdeI
MjclR	as	GCC C**AA GCT T**AT TTT TTA AAC TGT GTT CTA AGA ACT TTT TTA GGA GTT TGT AA	HindIII
AfAMAfll R	as	GCG C**GG ATC C**TT AAA CGA ACA TTT GCT GGG GCG GCT GTC TTG AGG G	BamHI
AfAM3A F	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TGT TAT CAC GCT GGA CGA TGT GGT TGG G	NdeI
CC-AMA R	as	CTC AAC TCA GCC GTT TCC CTC TTT GCC GGC GGC TGA CCG AGT CTG TCG	

		ACT TCT TC	
CC-AMA F	s	GTC GAC AGA CTC GGT CAG CCG CCG GCA AAG AGG GAA ACG GCT GAG TTG AGA TAC C	
Name	**Sense**	**Sequence (5'→ 3')**	**Comment**
AfAMAMF1	s	GGT GAA GGC CCG AAG GTC TTT GAT GTT TAC GCT AAA GAC CAG TGG AAG GGG GAG TTT GTG	
AfAMAMF2	s	GGA ATT CCA TAT GGC AAA GAG GGA AAC GGC TGA GTT GAG ATA CCT GAT AGT TCG TCC TTT GGG CTA CCC CCT CAA GGC GAG CTA TCA CGG TGA GGG CCC	
V-NC6F1	s	GAA TAT CCC CAG GTG GAT AAT ATT GAA GAA TAT GTT CAA CGT GCA CTG ATC CGT CGT CCG ATG CTG G	
V-NC6F2	s	GGA ATT CCA TAT GGT TAA AAA GGT TAC TCT GGC ACC GAT CAT TCG TAA GGA TCA GCG TCT GAA ATT TGG TGA GGG CGA ATA TCC CCA GGT GGA TAA T	
AfAMA Y-A	s	GCA AAG AGG GAA ACG GCT GAG TTG AGA TAC CTG ATA GTT CGT CCT TTG GGC GCA CCC CTC AAG GCG AGC TAT CAC GAA TAT CCC	
AfAMA L-A	s	GCA AAG AGG GAA ACG GCT GAG TTG AGA TAC CTG ATA GTT CGT CCT TTG GGC TAC CCC GCA AAG GCG AGC TAT CAC GAA TAT CCC	
AfAMA YL-A	s	GCA AAG AGG GAA ACG GCT GAG TTG AGA TAC CTG ATA GTT CGT CCT TTG GGC GCA CCC GCA AAG GCG AGC TAT CAC GAA TAT CCC	
AfAMA delta2	s	AAG GCG AGC TAT CAC GAA TAT CCC CAG GTG GAT AAT CCG	
AfAMA delta1	s	GCA AAG AGG GAA ACG GCT GAG TTG AGA TAC CTG ATA GTT CGT CCT TTG AAG GCG AGC TAT CAC GAA TAT CCC CAG G	
HMVNC_For	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TGC TCC CGT TGC GGA CCG	NdeI

Name	Sense	Sequence (5'→ 3')	Comment
		TAT CGT GCT CTC GTT TGC CCC CAG CGC CGC CGA TGG GGA C	
HMVNC_Rev	as	GCC C**AA GCT T**AT GGC GCG TCG GCG TCG TTA CGG GGA TGA ATG GCA AGC GTC GTT GCG TCG CC	HindIII
Name	**Sense**	**Sequence (5'→ 3')**	**Comment**
F1-AMAlo	s	AAG GAT CAG CGT CTG AAA TTT GGT GAA GGC AAG GTC TTT GAT GTT TAC GCT AAA GAC CAG TGG AAG GGG G	
F2- AMAlo	s	GGA ATT C**CA TAT G**GC AAA GAG GGA AAC GGC TGA GTT GAG ATA CCT GAT AGT TCG TAA GGA TCA GCG TCT GAA ATT TGG TGA AGG C	
For1-VatAMA	s	CCT TTG GGC TAC CCC CTC AAG GCG AGC TAT CAC GAA TAT CCC CAG GTG GAT AAT CCT ATT GAA GAA TAT GTT CAA CGT GCA CTG ATC CGT CGT CCG	
For2-VatAMA	s	GGA ATT C**CA TAT G**GT TAA AAA GGT TAC TCT GGC ACC GAT CAT TCG TCC TTT GGG CTA CCC CCT CAA GGC GAG CTA TCA CGA ATA TCC CCA GG	NdeI
VAMA F1	s	GGA ATT CCA TAT GCA TCA TCA TCA TCA TGT TAA AAA GGT TAC TCT GGC ACC GAT CAT TCG TCC TTT GGG CTA CCC CCT CAA GGC G	
VAMA R1	as	GCC TTC CTC TTT GCC TCT TCC TGC CCA ACC ACA TCG TCC AGC GTG ATG TCT CTG ACT TCC TCG AGA ACT TCT GAA GCC GG	
VAMA F2	s	CCG GCT TCA GAA GTT CTC GAG GAA GTC AGA GAC ATC ACG CTG GAC GAT GTG GTT GGG CAG GAA GAG GCA AAG AGG AAG GC	
AmconF	s	GGA AAG CGA CCC CAG AGT TAT CGA GAC CGA AAT TGT CAG AGA CAT CAC GCT GGA CGA TGT GGT TGG GCA GGA AGA GGC AAA GAG GAA GG	
FP1500	s	GGA ATT C**CA TAT G**GA GGG TGT CAT AAT GTC GGA GCT GAA GTT AAA GC	NdeI
AmconR	as	CCT TCC TCT TTG CCT CTT CCT GCC CAA CCA CAT CGT CCA GCG TGA TGT CTC TGA CAA TTT CGG TCT CGA TAA CTC TGG GGT CGC TTT CC	

FP1500	s	GGA ATT C**CA TAT G**GA GGG TGT CAT AAT GTC GGA GCT GAA GTT AAA GC	NdeI
FPHis1500	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TGA GGG TGT CAT AAT GTC GGA GCT GAA G	NdeI
Name	**Sense**	**Sequence (5'→ 3')**	**Comment**
RP1500	as	GCC C**AA GCT T**AC GTT CTG ATA AGG GTA AGT TTT TGT TCA TCA ATT ATC	HindIII
RP1500Xh	as	CCG **CTC GAG** TTA CGT TCT GAT AAG GGT AAG TTT TTG TTC ATC AAT TAT C	XhoI
RPh15AMA	as	GCC C**AA GCT T**AA GGA TGG GTA ACA AGC GTT ATT TTC GTC CTA TCT TCT ACC	HindIII
FPh15OB	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TGT TGA TGT ACT AGA GGC TAA AAT TAA GG	NdeI
CCPh15Rev	as	CAG GTT TTC ATC AGA GAT CAC ATC CTT AAT TCC TCC CGA CGG CGG CTG ACC GAG TCT GTC GAC TTC TTC ACG AAG TGC G	
Ph15CCFor	s	GTC GAC AGA CTC GGT CAG CCG CCG TCG GGA GGA ATT AAG GAT GTG ATC CTG GAT GAA AAC CTG ATA GTG GTA ATA ACT G	
EndoF	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TAA CAA AAA CTT ACC CTT ATC AGA ACG TGA AAG GGC CCT TAA GAT AAT AAA GAT TCT C	NdeI
EndoR	as	GGC C**AA GCT T**AT TGG CTA GAG GTA TCC TGA ACG CCA ATT TTT GGA CAG AGT CCC C	HindIII
HVVncF	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TAC GAC CGA ACC GAC GGC GAC CCG CGT GGT CGT CTC C	NdeI
HVVncR	as	GCC C**AA GCT T**AC CGCT GTT CCG CCG TCG AGG AGA CCG GCC CGT CCG CGC TCT GGA CGC	HindIII
CC-F	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TAG CTC GAC AGA GAA CCC	NdeI

Name	Sense	Sequence (5'→ 3')	Comment
		GGA TTC GG	
CC-VNC R	as	AAC CTT TTT AGC GAT TTC GGT ACG CGG CGG CTG ACC GAG TCT GTC GAC TTC TTC	
CC-VNC F	s	GAA GTC GAC AGA CTC GGT CAG CCG CCG CGT ACC GAA ATC GCT AAA AAG GTT ACT CTG GC	
AfNCthioF		CAT G**CC ATG G**CA AAG AGG GAA ACG GCT GAG TTG AGA TAC CTG ATA GTT CGT CCT TTG GG	NcoI
AfNCthioR		GC**T CTA GA**G CTT AGT CTC TGA CAA TTT CGG TCT CGA TAA CTC TGG GG	Xba1
PAN and ARC related project primers			
ForPyr	s	GGG AAT TC**C ATA TG**C ATC ATC ATC ATC ATC ATG TGA GTG TAA TGA GTG GTG ACG AAG	NdeI
RevPyr	as	CGC CC**A AGC TT**A TTC TAT GAC CTC AAA TCC CAG TAC TG	HindIII
Arch fulg F	s	GGG AAT TC**C ATA TG**C ATC ATC ATC ATC ATC ATG GCG ATA GCG AAA TAC AAT ACC TCC	NdeI
Arch fulg R	as	CGC CC**A AGC TT**A CTC CTC CAC CTC AAA GCC GTA AAC C	HindIII
Met. Jan. F	s	GGG AAT TC**C ATA TG**C ATC ATC ATC ATC ATC ATG TTT TTG AAG AAT TTA TTT C	NdeI
Met. Jan. R	as	CGC CC**A AGC TT**A TTC ATC AAC TTC CAT TGC TTT AGC	HindIII
MjPNG F	s	CGC **GGA TCC** GTT TTT GAA GAA TTT ATT TCA ACT GAA TTG AAG	BamHI
APaFll R	as	`CCG` **`CTC GAG`** `TTA AAC GAA CAT CAC TCC CTT CAA ATC GGG TAT CGG AGT CGT CTT CTT` `GAG C`	XhoI

Name	Sense	Sequence (5'→ 3')	Comment
AfPthioF	s	CAT G**CC ATG G**GC GAT AGC GAA ATA CAA TAC CTC CTG GAA AAG TTG AAG AAG CTG G	NcoI
AfPthioR	as	GC**T CTA GA**G CTT ACT CCT CCA CCT CAA AGC CGT AAA CCA TTG GGT CTT TGG	XbaI
APccF	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TCC GCC ATT ACT CGT AGG TGT TGT TTC	NdeI
CC forw	s	CGA AGA TCT GGT TCT CGA GGA ACC GCC GAG CTC GAC AGA GAA CCC GGA TTC GGT TGC GG	
CC rev	as	GCC C**AA GCT T**AC TGA CCG AGT CTG TCG ACT TCT TCA CGA AGT GCG ATG AGC TGC	HindIII
ID forw	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TCC GCC GAG CGG TTA CGG AGT CCT TCT C	NdeI
ID rev	as	GGT TCT CTG TCG AGC TCG GCG GTT CCT CGA GAA CCA GAT CTT CGA CCT CGG CCT TCG G	
OB1 rev	as	GCC CAA GCT TAC TGC TCG TAG GTG CCG GCC TCG ACG ATG GTG AGT GCC	
Farcpancc	s	GGC TGA GAA GTG AAG TTG AAA GAT TAC GCT CAC CGC CGA GCG GTT ACG GAG TCC TTC TCT CGG TAC ACG AGG	
Rarcpancc1	as	CCT CGT GTA CCG AGA GAA GGA CTC CGT AAC CGC TCG GCG GTG AGC GTA ATC TTT CAA CTT CAC TTC TCA GCC	
Rarcpancc2	as	CCG **CTC GAG** TTA TTC CTC GAG AAC CAG ATC TTC GAC CTC GGC CTT CGG GAT GCG CTC	XhoI

Name	Sense	Sequence (5'→ 3')	Comment
Fpanarccc	s	CGT GAA GAA GTC GAC AGA CTC GGT CAG CCG CCG TTA CTC GTA GGT GTT GTT TCA GAC ATA CTT GAA GAC GG	
Rpanarccc	as	CCG TCT TCA AGT ATG TCT GAA ACA ACA CCT ACG AGT AAC GGC GGC TGA CCG AGT CTG TCG ACT TCT TCA CG	
FullNosFw	s	GGG AAT TC**C ATA TG**C ATC ATC ATC ATC ATC ATG ACG ATA TTG ATT TAC AGA ACC TTG ACA GTT GG	NdeI
NosRev	as	CGC CC**A AGC TT**A CCC ACC ACT ACG AAT ATG TAG GCG AGG CTT AAC TTT AGC	HindIII
NosOBFw	s	GGG AAT TC**C ATA TG**C ATC ATC ATC ATC ATC ATT TGG AAT ATG GGG TCT TTC TCG TAA AT GCC C	NdeI
GcnAfPanF1	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TAG AAT GAA ACA ACT TGA AGA CAA GGT TGA AGA ATT GCT TTC GAA AAA TTA TCA CTT GG	NdeI
GcnAfPanF2	s	GTT GAA GAA TTG CTT TCG AAA AAT TAT CAC TTG AAA AT GAG GTT GCC AGA TTA AGA TCG CCG CCA TTA CTC GTA GGT GTT GTT TCA GAC ATA CTT G	
GAPP-AAF2	s	GTT GAA GAA TTG CTT TCG AAA AAT TAT CAC TTG AAA AT GAG GTT GCC AGA TTA AGA TCG GCC GCC TTA CTC GTA GGT GTT GTT TCA GAC ATA CTT G	
GAPP-PAF2	s	GTT GAA GAA TTG CTT TCG AAA AAT TAT CAC TTG AAA AT GAG GTT GCC AGA TTA AGA TCG CCG GCC TTA CTC GTA GGT GTT GTT TCA GAC ATA CTT G	
GAPP-PWF2	s	GTT GAA GAA TTG CTT TCG AAA AAT TAT CAC TTG AAA AT GAG GTT GCC AGA TTA AGA TCG CCG TGG TTA CTC GTA GGT GTT GTT TCA GAC ATA CTT G	
GAPP-APF2	s	GTT GAA GAA TTG CTT TCG AAA AAT TAT CAC TTG AAA AT GAG GTT GCC AGA TTA AGA TCG GCC CCA TTA CTC GTA GGT GTT GTT TCA GAC ATA CTT G	

Name	Sense	Sequence (5'→ 3')	Comment
GCARCF2	s	GTT GAA GAA TTG CTT TCG AAA AAT TAT CAC TTG GAA AAT GAG GTT GCC AGA TTA AGA TCG CCG CCG AGC GGT TAC GGA GTC CTT CTC TCG G	
MjchArR1	as	AGC CAA TAC TTG TCC TTT TCT TCT GTC AAT TAT TGG CGG CTG ACC GAG TCT GTC GAC TTC TTC ACG AAG TGC G	
MjchArF2	s	CGC ACT TCG TGA AGA AGT CGA CAG ACT CGG TCA GCC GCC AAT AAT TGA CAG AAG AAA AGG ACA AGT ATT GGC T	
MjchArR2	as	GCC C**AA GCT T**AC TTT CCA CCA ATA ACT CTT GTT ATC TTG TAT TGA CCA ACT GC	HindIII
MjIAr F2	s	CGT GAA GAA GTC GAC AGA CTC GGT CAG CCG CCA AGC GGA CAA GTA TTG GCT ATA ATG GGA GAT ATG GTT CAA ATT ATG GAC TTG CAA ACT TAC G	
MjIArR1	as	CGT AAG TTT GCA AGT CCA TAA TTT GAA CCA TAT CTC CCA TTA TAG CCA ATA CTT GTC CGC TTG GCG GCT GAC CGA GTC TGT CGA CTT CTT CAC G	
EcchArR	as	GCG TTG AAC CAT TTT ACG ATA CCA GTC ATT TTA CCG GAC AGC AGT GGC GGC TGA CCG AGT CTG TCG ACT TCT TCA CGA GTG CG	
EcchArF1	s	CGC ACT TCG TGA AGA AGT CGA CAG ACT CGG TCA GCC GCC ACT GCT GTC CGG TAA AAT GAC TGG TAT CGT AAA ATG GTT CAA CGC	
EcchArR1	as	GCC C**AA GCT T**AC AGG CTG GTT ACG TTA CCA GCT GCC GGG CCT TTA GCG CCG C	HindIII
EccSAR	as	GCC CAA GCT TAG TAG GTG CCG GCC TCG ACG ATG GTG AGT GCC TCG TTG AGG CGC ACC GTC TGA CCT TCG TCC AGA GAT TTG TAA CCA TCG	
GCECCSF2	s	GTT GAA GAA TTG CTT TCG AAA AAT TAT CAC TTG GAA AAT GAG GTT GCC AGA TTA AGA TCG CCG CC ACT GCT GTC CGG TAA AAT GAC TGG TAT CG	
Name	**Sense**	**Sequence (5'→ 3')**	**Comment**
EccFor	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TTC CGG TAA AAT GAC TGG	

Name	Sense	Sequence (5'→ 3')	Comment
		TAT CGT AAA ATG G	
GCECANwF2	s	GTT GAA GAA TTG CTT TCG AAA AAT TAT CAC TTG GAA AAT GAG GTT GCC AGA TTA AGA TCG CCG CCA TCC GGT AAA ATG ACT GGT ATC GTA AAA TGG	
EcchArNewR	as	GCG TTG AAC CAT TTT ACG ATA CCA GTC ATT TTA CCG GAT GGC GGC TGA CCG AGT CTG TCG ACT TCT TCA CGA AGT GCG	
EcchArNewF1	s	CGC ACT TCG TGA AGA AGT CGA CAG ACT CGG TCA GCC GCC ATC GGT AAA AAT GAC TGG TAT CGT AAA ATG GTT CAA CGC	
AbrPaR	as	CGT AAT CTT TCA ACT TCA CTT CTC AGC CTT CTT ACT TCT CTT TCA TAG CTG CCG TTT GGT TTA TAT TTT TTA AGG ATG ATT TTT TCA TC	
ABrPaF	s	GAT GAA AAA ATC ATC CTT AAA AAA TAT AAA CCA AAC GGC AGC TAT GAA AGA GAA GTA AGA AGG CTG AGA AGT GAA GTT GAA AGA TTA CG	
Arc2OBF	s	GGA ATT C**CA TAT G**CA TCA TCA TCA TCA TCA TGT CGG TGA GAT CAG CAC GCT CCG TGA GGT ACT CGA CGA CGG C	NdeI
Arc2OBR	as	GCC C**AA GCT T**AT TCC TCG AGA ACC AGA TCT TCG ACC TCG GCC TTC GGG ATG CGC TCG AAG G	HindIII
ARCPP-AAF	s	GCA CTT CGT GAA GAA GTC GAC AGA CTC GGT CAG GCC GCC AGC GGT TAC GGA GTC CTT CTC TCG GTA CAC GAG G	
ARCPP-AAR	as	CCT CGT GTA CCG AGA GAA GGA CTC CGT AAC CGC TGG CGG CCT GAC CGA GTC TGT CGA CTT CTT CAC GAA GTG C	
ARCPP-PAF	s	GCA CTT CGT GAA GAA GTC GAC AGA CTC GGT CAG CCG GCC AGC GGT TAC GGA GTC CTT CTC TCG GTA CAC GAG G	
Name	**Sense**	**Sequence (5'→ 3')**	**Comment**
ARCPP-PAR	as	CCT CGT GTA CCG AGA GAA GGA CTC CGT AAC CGC TGG CCG GCT GAC CGA GTC TGT CGA CTT CTT CAC GAA GTG C	

ARCPP-PWF	s	GCA CTT CGT GAA GAA GTC GAC AGA CTC GGT CAG CCG TGG AGC GGT TAC GGA GTC CTT CTC TCG GTA CAC GAG G	
ARCPP- PWR	as	CCT CGT GTA CCG AGA GAA GGA CTC CGT AAC CGC TCC ACG GCT GAC CGA GTC TGT CGA CTT CTT CAC GAA GTG C	
ARCPP-APF	s	GCA CTT CGT GAA GAA GTC GAC AGA CTC GGT CAG GCC CCG AGC GGT TAC GGA GTC CTT CTC TCG GTA CAC GAG G	
ARCPP-APR	as	CCT CGT GTA CCG AGA GAA GGA CTC CGT AAC CGC TCG GGG CCT GAC CGA GTC TGT CGA CTT CTT CAC GAA GTG C	

3. Methods

3.1 Molecular biology methods

3.1.1 Polymerase chain reaction (PCR)

The polymerase chain reaction was carried out with specific primers to amplify DNA fragments from plasmids, genomic DNA or primers themselves. The same method was used to introduce restriction sites, deletions, insertions and point mutations. Annealing temperatures (Ta) for primers were calculated with the program "Tm of primers" (made by Yakov Sergeev), and if they were different for each primer the lower one was chosen.

Total volume of the reaction was usually 50μl, except in colony PCRs where it was 30 μl. A reaction usually contained up to 2ng of plasmid or 500ng of genomic DNA, 100μM dNTPs, 0.6-1pM of primers, 1U Pfu-Polymerase (in case of colony-PCR 0.5U of Taq-Polymerase was used) and 1x of the recommended reaction buffer.

The temperature program is shown in Table 3.1. Length of cycles was varying for different constructs depending on the length of DNA fragments that were amplified.

Denaturation		94°C
18-27 cycles	Denaturation	94°C
	Annealing	*Ta*
	Extension	72°C
Final extension		72°C

Table 3.1 Typical program used for PCR reaction

3.1.2 Isolation and purification of DNA

Plasmid DNA was isolated from bacterial cultures by a small-scale purification method, known as mini-prep. The standard protocol of the QIAGEN Miniprep Kit with centrifuge was carried out using 1-3ml of 6-8 hours old bacterial culture, giving a yield of 10-20ng DNA. DNA was eluted with sterile water to be used directly for DNA sequencing and restriction.

To check yield, purity, size and digestion of DNA plasmids and amplified fragments, agarose electrophoresis was carried out. Agarose was dissolved in TAE-buffer using a microwave oven. To detect DNA, ethidium-bromide was added to a final concentration of 0.5μg/ml in the cooled solution. The solution was poured in a mould containing a well-forming comb.

The agarose content was chosen according to the expected size of DNA-fragments:

$$\geq 2000\text{bp} \qquad 0.8 - 1\,\%$$

$$500 - 2000\text{bp} \qquad 1 - 1.5\,\%$$

$$\leq 500\text{bp} \qquad 1.5 - 2\,\%$$

TAE-buffer was used for electrophoresis. Agarose gels were submerged in a horizontal electrophoresis apparatus and DNA samples mixed with loading buffer were loaded. Electrophoresis was performed at 60-90V for 30-60 min at room temperature. DNA size markers (Fermentas GeneRuler^{TM} 50bp or 1kb DNA ladder) were included, depending on size of electrophoresed DNA fragments (Appendix). For detection and photographing of DNA the Biorad Chemi Doc™ gel documentation system was used with the Quantity One® software.

DNA fragments separated by electrophoresis were excised from agarose gels and extracted with the Qiagen Gel extracting Kit following the manufacturer's protocol. DNA was eluted with sterile water and nucleic acid content was measured by absorbance at 260

nm. An OD of 1 corresponds to 50µg/ml for double stranded DNA. The ratio OD_{260}/OD_{280} was calculated to estimate purity of DNA (ratio of 1.8 was desirable).

3.1.3 Photometric determination of DNA concentration

Nucleic acid concentration was measured at 260 nm. Nucleic acid concentration was calculated using the following equation: c [µg/µl] = E260 * f * (dilution), f = 0.02 (for oligonucleotides, ss-DNA, RNA), f = 0.04 (for ds-DNA, plasmids). The ratio OD_{260}/OD_{280} was calculated to estimate purity of DNA. An OD ratio of >1.8 was desirable.

3.1.4 DNA digestion with restriction enzymes

Double-stranded DNA molecules were digested with 2-5 U of the restriction enzyme per µg of DNA at optimal temperature (usually 37°C) in the buffer recommended by the supplier. If digestion was done with two enzymes, the most compatible buffer was used. In cases where it was not possible to do digestion with two or more enzymes at the same time, sequential restriction was used with DNA electrophoresis and purification steps in-between(Qiagen Gel extracting Kit) .

3.1.5 5'-DNA-dephosphorylation

To avoid re-ligation of blunt-ended plasmid DNA after digestion, the 5´-phosphates were removed with calf alkaline phosphatase. 1-2 µg DNA were incubated for 30 min at 37°C with 0.2U alkaline phosphatase in phosphatase buffer. Dephosphorylated plasmid DNA was loaded on a DNA agarose gel, electrophoresed and purified by standard procedure with the Qiagen Gel extracting Kit.

3.1.6 Ligation of DNA fragments

DNA fragments were ligated with DNA ligase, using the T4 DNA Ligation Kit (NEB or Fermentas) according to supplier instructions. If possible, a vector/insert ratio 1:5 to 1:9, was used.

3.1.7 Purification of DNA by precipitation with ethanol

To precipitate DNA from solution, 1/10 of the total volume 3 M Na-acetate (pH 4.5) and two volumes of 98.9% ice-cold ethanol were added. The sample was precipitated at -20°C for 1 h. The sample was centrifuged for 15 min at 12000 × g and 4°C and the supernatant was discarded. The pellet was washed with 750 µl of chilled 70 % ethanol and centrifuged for 10 min at 12000 × g and 4°C. The supernatant was again discarded, and the pellet was dried and re-suspended in appropriate 1 x buffer or sterile dH_2O.

3.1.8 DNA sequencing

DNA sequencing was done in automated "in house" genome sequencing facility, using fluorescent dye labelling, by the following program:

	Temperature (°C)	Time
	96	20''
30 cycles	50	20''
	40	4'
	4	∞

Mixture for the PCR reaction contained 5-10pM of sequencing primers, BDT mix (**B**ig **D**ye **T**erminator), 1x sequencing buffer and 10-20pM of plasmid DNA in final volume of 10µl.

3.2 Cloning strategies

All constructs were cloned directly into the expression vectors and positive clones were sequenced. Vectors that were used for expression were pET30b, pGEX4T-1 and pThioHis (Fig. 3.1), with exception of ARC-N, ARC-Nn and ARC-Nc clones in the pT7-7 vector, which were a generous gift of Dr. Peter Zwickl from the MPI for Biochemistry in Martinsried. Constructs in pGEX4T-1 and pThioHis were N-terminal fused to glutathione S-transferase (GST) and thioredoxin, respectively.

The Hind III site which does not exist in the commercial pGEX4T-1 vector was introduced with the primer of one clone.
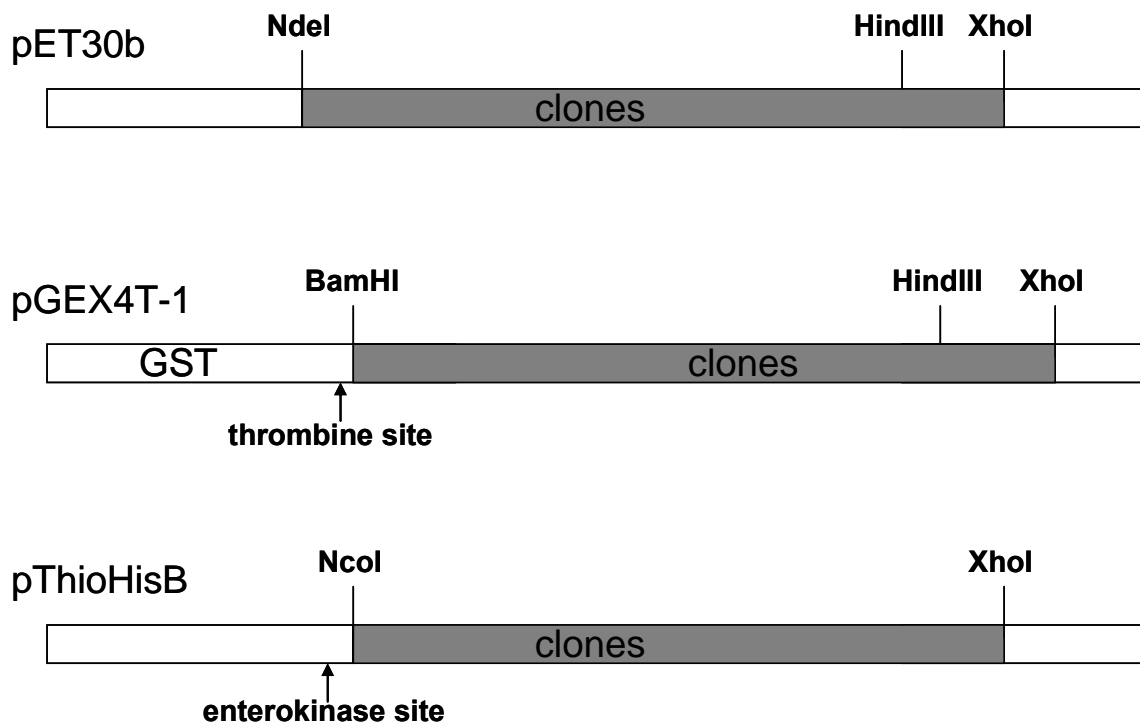

Figure 3.1 Cloning strategy for pET30b, pGEX4T-1 and pThioHis vectors that was used for most of the constructs

Chimeric proteins were made using overlapping primers (Fig. 3.2 A) and without use of the additional restriction sites which simplified cloning strategy to the one shown in Fig. 3.1. The same strategy was used for the clones with mutations in the connecting regions or inside the genes (Fig. 3.2 B).
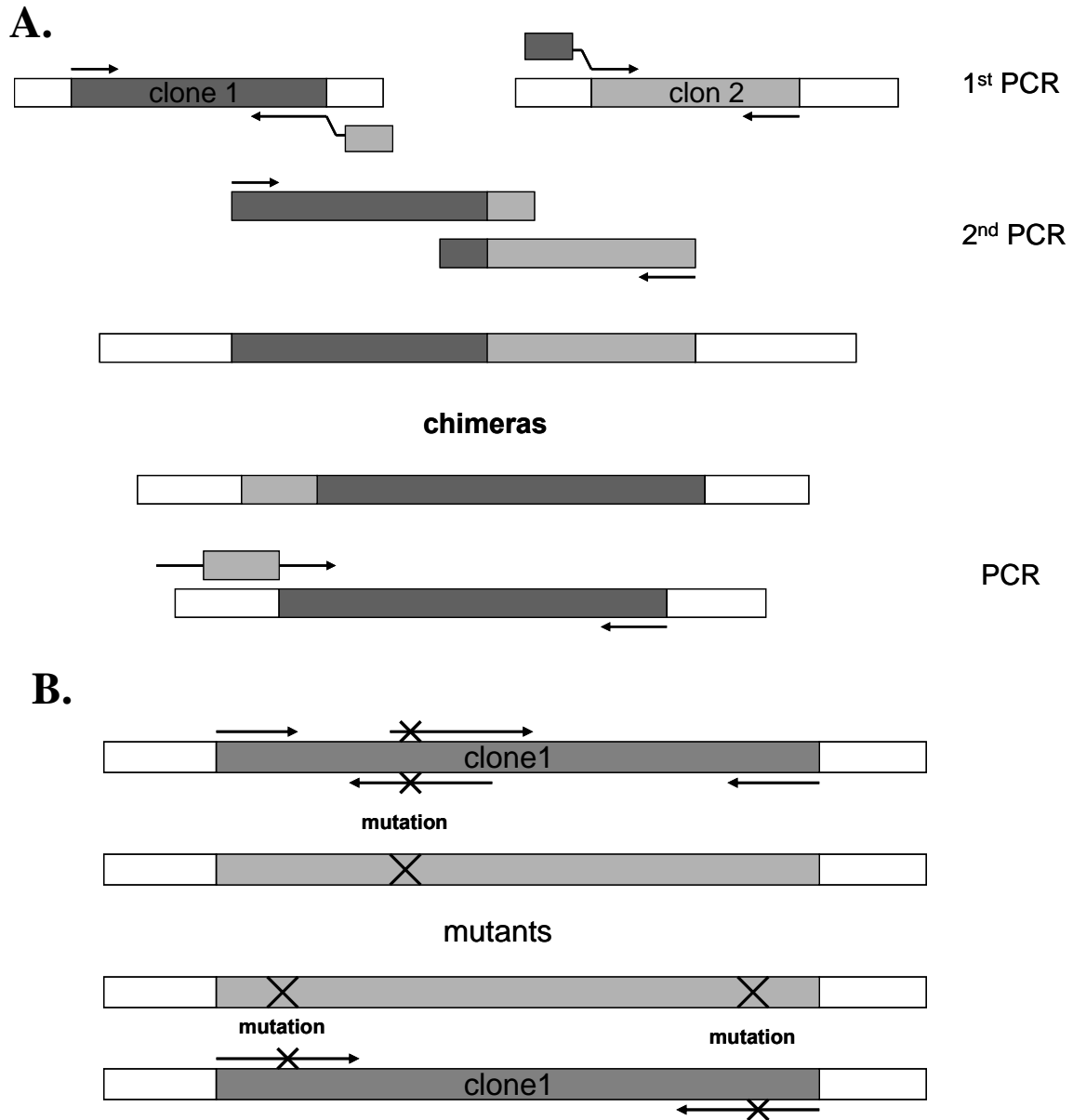
A.

B.

Figure 3.2 (A.) Simplified schemes for producing chimeras and **(B.)** mutants

A more complicated strategy was used to produce duplicated constructs (VAT-Nnc db, MT6002757 db, etc), as it involved introduction of the ribosome binding site (Rbs) and restriction sites in the primers that were used (Fig. 3.3). Some of the constructs were built entirely from primers by serial PCR reactions, and the final sequence was introduced directly into the expression vector like in Figure 3.1.
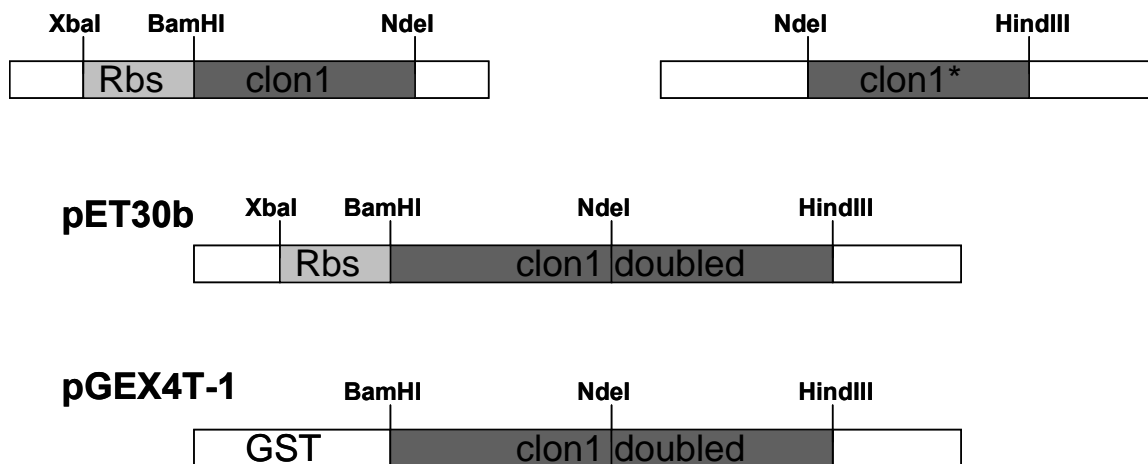

Figure 3.3 Simplified scheme for making doubled constructs and their insertion in expression vectors

3.3 Microbiological methods

3.3.1 Competent cells

5 ml of LB medium were inoculated with a single bacterial colony (TOP10, BL21, BL21Gold or C41 cells) and incubated over night at 37°C at 200 rpm. 0.5 ml of the culture were transferred into 50 ml of fresh LB medium and grown at 37°C (200 rpm) to 0.3-0.4 OD_{600}. The culture was cooled on ice for 10 min and centrifuged at 2500 × g for 10-15 min at 4°C. The pellet was resuspended in 10 ml ice cold $CaCl_2$ (0.1 M) and centrifuged once more (2500 × g, 10 min, 4°C). The pellet was again resuspended in 2 ml

ice cold 0.1 M CaCl$_2$ supplemented with 15-20 % glycerol and incubated at least for 2 h on ice, aliquoted (100 µl) and stored at -80°C.

3.3.2 Standard transformation of competent E. coli cells

The DNA ligation reaction (10 µl) was added to 100 µl competent cells, mixed gently and then incubated on ice for 30 min. Cells were heat-shocked at 42°C for 1 minute and incubated on ice for 10 min. 600 µl of LB medium were added and cells were incubated at 37°C for 45 min with agitation (200-250 rpm). Cell suspension was centrifuged shortly (5 seconds) and 600 µl of the LB medium were removed. The remaining 100 µl were spread on LB agar plates with appropriate antibiotics. Plates were incubated in an inverted position for 12-16 h at 37°C.

3.3.3 Rapid transformation of competent E. coli cells

One aliquot of competent cells was thawed on ice. 100 ng of plasmid DNA (max) was added and mixed. After 20 minutes on ice the mixture was spread on LB agar plates containing the appropriate antibiotic and incubated at 37°C O/N.

3.3.4 Glycerol stock cultures

1-3 ml of an over night bacterial culture was centrifuged (2 min, 10,000 x g). The pellet was suspended in 1 ml of a LB medium containing 20% glycerol and stored at -80°C. Constructs for the clone libraries were kept in Top10 cells to avoid recombination events that can occur in some of the expression strains.

3.4 Protein chemistry methods

All constructs were expressed in the *E. coli* BL21 gold (DE3) or C41 (DE3) cells. Constructs were tested for expression in a small volume (5ml) prior to expression on large scale either in LB medium or in a minimal medium with the labeled compounds and with corresponding antibiotics. 800μl of the non-induced culture were mixed with 200μl of the 80% glycerol and kept as glycerol stocks in a -80°C freezer. Fresh transformants or glycerol stocks were used for inoculation of overnight pre-cultures. Such pre-cultures were used for the inoculation of a big-scale expression in the amount of 2-3% of the final medium volume. Cells were usually grown at 37°C, induced at $OD_{600}\sim0.6$ with 1mM IPTG and harvested after 4h.

3.4.1 Expression of labeled proteins for NMR spectroscopy

Comprehensive study of protein structure by nuclear magnetic resonance (NMR) spectroscopy would not be possible without labeling of the recombinant proteins of interest (Goto and Kay, 2000). Bacterial cultures were grown in M9-minimal medium consisting of: M9-salt mixture, glucose, NH_4Cl, $MgSO_4$, $CaCl_2$, vitamin supplements, micronutrients and adequate antibiotic(s). After adjusting optimal conditions for the protein expression, pre-cultures were grown in medium containing labeled compound(s), ^{15}N-labeled NH_4Cl (0.7g/l) and/or ^{13}C-labeled glucose (2g/l). In the cases where triple labeling was needed (Ph1500C and Ph1500), cells were first grown in increasing amounts of D_2O in medium, to get accustomed to it. When 80% of D_2O was reached in the medium, pre-cultures were grown containing also deuterated-^{13}C-glucose ($^{2}H^{13}C$-glucose) and ^{15}N-labeled NH_4Cl. Use of the $^{2}H^{13}C$-glucose enable more then 90% deuteration of protein side chain hydrogen atoms. Suppression of the ^{1}H-^{1}H relaxations plays a great role in obtaining distances and constrains in large proteins or multimers like in our case (Le Master, 1990). Cells were grown at 37°C, induced at $OD_{600}\sim0.6$ and

harvested after 4 to 6 hours. Cell cultures for the triple labeled samples where induced and harvested after overnight expression at 30°C.

3.4.2 Expression of Se-Met labeled proteins for crystallography

Plasmids expressing proteins of interest were first transformed to the Met-auxotroph strain of *E. coli*, B834 (DE3) (Novagen). Expression of the proteins in this strain gives 99% labeling of Met-residues. Test expression was done from fresh transformants in M9-minimal medium with the double amount of the M9-salts and glucose (4mg/l), and unlabeled Met (4mg/l). After optimizing expression conditions cells were grown in medium containing the same amount (4mg/l) of Se-Met instead of the unlabeled Met. Cells were grown at 37°C, induced at OD_{600}~0.8 and harvested after overnight expression at 30°C.

3.4.3 Purification of soluble proteins

3.4.3.1 Proteins with 6xHis-tag or thioredoxin fusions

Frozen cell pellets containing protein with either 6xHis-tagg or N-terminally fused thioredoxin, modified to bind metal chelating resins, were suspended in lysis buffer (A) and thawed on ice (15 min). Prior to lysis by French press, 1 mM MgCl2, 0.1 mM PMSF and DNAse (Sigma) were added to mixture. Cells were additionally lysed by sonification (microtip setting 5, on ice, 3 x 10 s) in case of 6xHis tagged proteins. Cell homogenate was centrifuged (16000 x g, 30 min, 4°C). Supernatant was filtrated through a sterile filter (Millipore, ϕ0.22 μm) and applied to a self-packed column with Ni-NTA Superflow resin (GE-Healthcare). Purification is done on the HP Pharmacia Purifier or ÄKTA Purifier (GE Healthcare). Column was washed with washing buffer (B) until OD_{280} was constant. The protein of interest was eluted in 5 ml fractions with a linear gradient from

50 mM (B) to 1M imidazole elution buffer (C). Fractions were checked by SDS-PAGE and those containing protein of interest were pooled and dialyzed overnight against the buffer that was used for gel-filtration. The protein sample was concentrated with Vivaspin or Millipore centricons, with different cut-off membranes depending on the protein size, and applied on a gel-filtration column (superdex 75, superdex 200 or whatever; GE healthcare, depending on protein size). Fractions containing protein were pooled, concentrated and kept at 4°C with addition of Na-azide (final conc. 0.02% v/v) or at -80°C in buffer containing glycerol (final v/v 15%) before usage.

3.4.3.2 Proteins with GST-tag

A frozen cell pellet containing protein with an N-terminal fusion to GST were suspended in PBS buffer and thawed on ice (15 min). $MgCl_2$ (1 mM), PMSF (0.1 mM) and DNAse (Sigma) were added prior to lysis by French press. Cell homogenate was centrifuged (16000 x g, 30 min, 4°C). Supernatant was filtrated through a sterile filter (Millipore, ϕ0.22 µm), applied on a pre-packed 5 ml GSH column (GE-Healthcare) and washed with PBST (PBS buffer + 0.1% Triton X-100) buffer until OD_{280} reached a baseline. Protein was eluted in 2.5 ml fractions with GST-elution buffer (D) and analyzed by SDS-PAGE. Fractions containing fusion protein were pooled and concentrated. 30xThrombin cleavage buffer (TCB), 2-mercaptoethanol (βME) (final conc. 0.1% v/v) and thrombin (2U per mg of fusion protein) were added to the sample and mixed 2 hours at room temperature (or overnight at 4°C) to perform thrombin cleavage. The cleaved sample was dialyzed against PBST buffer and applied to the GST column again. The column flowthrough was collected, concentrated and checked for protein purity by SDS PAGE. Protein samples were filtrated (Millipore, ϕ0.22 µm) and applied on the gel filtration column. Fractions containing protein were pooled, concentrated and stored at 4°C with addition of Na-azide (final conc. 0.02% v/v) or in buffer containing glycerol (final v/v 15%) at -80°C.

3.4.3.3 Proteins without tags

The frozen cell pellet containing the protein of interest was thawed on ice in low salt buffer (pH 7.0-7.5, 50 mM NaCl or KCl). Prior to lysis by French press, 1 mM MgCl$_2$, 0.1 mM PMSF and DNAse (Sigma) were added to mixture. The cell homogenate was centrifuged (16000 x g, 30 min, 4°C). Supernatant was filtrated through a sterile filter (Millipore, ϕ0.22 μm) and applied on a high load anion or cation exchange chromatography column, depending on the protein's isoelectric point. Proteins were eluted in 5-8 ml fractions by a linear gradient with high salt buffer (pH 7.0-7.5, 1 M NaCl or KCl). Fractions were analyzed by SDS-PAGE and those containing the protein of interest were pooled, concentrated and, if protein was not pure enough, dialyzed against low-salt buffer for a re-run on the ion exchange chromatography. The protein sample was then applied on a gel filtration column for the final purification step.. Fractions containing protein were pooled, concentrated and kept at 4°C with addition of Na-azide (final conc. 0.02% v/v) or in the buffer containing glycerol (final v/v 15%) at -80°C till used .

3.4.4 Purification of the insoluble proteins

3.4.4.1 Purification of the insoluble proteins (6xHis-tag)

Frozen cell pellet containing protein of interest were suspended in lysis buffer (A) and thawed on ice (15 min). Prior to lysis by French press 1 mM MgCl2, 0.1 mM PMSF and DNAse (Sigma) were added to mixture. Cell homogenate was centrifuged (16000 x g, 30 min, 4°C). Loading buffer containing 8 M urea (A1) was added to the pellet and incubated shaking at room temperature for 30 min. The mixture was centrifuged (16000 x g, 30 min, 20°C) and the supernatant was filtrated (Millipore, ϕ0.22 μm). The filtrated supernatant was applied on a Ni-NTA column equilibrated with the loading buffer (A1). The column was washed with loading buffer (A1) and washing buffer (B1) until OD$_{280}$

reached baseline. The protein was eluted either by elution buffer (C1) or washing buffer (B1) containing 0.5 M imidazole, and 5ml fractions were collected. Fractions containing protein were pooled and dialysed over night against refolding buffer. The usual refolding buffer contained 30 mM TRIS/HCl or phosphate buffer, 150 mM salt (NaCl or KCl) and 5% glycerol, but this composition was changed depending on the protein's behavior. In some cases, denatured protein was refolded by rapid dilution in refolding buffer. The refolded protein was concentrated and applied to a gel filtration column for the last step of the purification. Fractions containing protein were pooled, concentrated and kept at 4°C with addition of Na-azide (final conc. 0.02% v/v) or in the buffer containing glycerol (final v/v 15%) at -80°C till used.

3.4.4.2 Purification of insoluble non-tagged proteins

Frozen cell pellet containing protein of interest were suspended in a buffer with low salt (50mM TRIS/HCl pH 7.2, 50mM NaCl) and thawed on ice (15 min). Prior to lysis by French press 1 mM MgCl2, 0.1 mM PMSF and DNAse (Sigma) were added to mixture. Cell homogenate was centrifuged (16000 x g, 30 min, 4°C). Protein of interest that was in inclusion bodies was resolubilized by addition of low salt buffer containing 8 M urea. After 30 minutes of shaking on room temperature, the mixture was centrifuged (16000 x g, 30 min, 20°C) and the supernatant was filtrated (Millipore, ϕ0.22 μm). Filtrated supernatant was loaded on an anion or cation exchange chromatography column and the column was washed until OD_{280} reached baseline level. Protein was `eluted in` fractions of 5-10 ml by a linear gradient of high salt buffer (50 mM TRIS/HCl pH 7.2, 1 M NaCl) containing 8 M urea. Fractions containing protein were pooled and, if additional purification was needed, run over a second ion-exchange column. Refolding and storage of the protein was done like in case of his-tagged proteins (chapter 3.4.4.1)

3.4.5 Circular dichroism (CD) spetroscopy and measurment of thermal stability

Circular dichroism spectroscopy was used to estimate the secondary structure content of the protein. All CD spectra were recorded at room temperature at 190-240nm (6-fold accumulation) with a JASCO J-810 Spectropolarimeter, using 1 mm cuvettes at 0.2 nm resolution, 1nm bandwidth, 1 s time constant and sensitivity of 100 mdeg. For the measurements of CD spectra, protein samples (1-10 µM) were dialyzed against PBS or 10 mM TRIS/HCl pH 7.2-7.6, 150 mM NaCl buffer. When it was possible lower salt (or NaF) and borate buffer were used to overcome the problem of the absorption of circularly polarized light by some ions.

Thermally induced protein denaturation was also monitored by CD spectroscopy using a Peltier-controlled sample holder unit. In this case 1 cm cuvettes were used and protein sample was mixed with a magnet. CD spectra were recorded for the protein at room temperature and 95°C prior to measurement of the thermal denaturation curve. The wavelengths which showed the biggest changes between the two different temperatures were used to determine the melting temperature of the protein. These were usually 222 nm and 216nm, where the changes in α-helical and β-strand signals are highest. Temperature profiles were recorded in 1°C increments with 0.2° pitch from room temperature to max 98°C. In all cases, a temperature probe connected to the cuvette was used to provide an accurate temperature record.

3.4.6 Protein concentration

Protein concentration during purification steps was determined using three different methods. For most of the pure protein samples, ultraviolet absorption of the proteins at 280 nm was used. Extinction coefficients for proteins were calculated using ProtParam software (http://www.expasy.ch/tools/protparam.html - Gasteiger et al., 2005) and protein concentration was calculated from the measured absorbance.

Concentration of the protein was double checked using a commercial available BCA assay (Pierce; Smith et al., 1985). Protein samples were diluted 5-10 times in a final volume of 50 µl in the cuvettes for measurement, and subsequently 1 ml of BCA working

solution was added. Serial dilutions of a 1 mg/ml BSA stock were used for preparing a standard curve. Mixtures were incubated 30 minutes at 37°C until the color reaction was developed. Samples were cooled to room temperature and absorbance at 562 nm was determined. Protein concentration was determined from the standard curve equation.

A modification of the Bradford assay (Bradford, 1976), the "drop assay", was used to follow purification of the protein on a gel filtration column or in cases when protein did not absorb at 280 nm. 10 µl of the fractions from gel filtration were mixed with 30 µl of the Bradford reagent in a drop on a clean stripe of Parafilm® M. Steady blue color was a sign for the protein presence in tested fractions, which were then further analyzed by SDS-PAGE.

3.4.7 SDS-polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-PAGE electrophoresis was done using the GE-Healthcare apparatus. In table 3.2 the recipe for polyacrylamide gels are shown:

Stacking gel	4 %	Running gel	18 %	15 %
Stacking gel buffer	9.4 ml	Running gel buffer	18.75 ml	18.75 ml
Water	18.4 ml	Water	3 ml	10.5 ml
AA/Bis (37.5%):(1%)	5.6 ml	AA/Bis 37.5:1	45 ml	37.5 ml
TEMED	1.9 ml	TEMED	3.75 ml	3.75 ml
APS 1.4%	1.9 ml	APS 1.4%	3.75 ml	3.75 ml
SDS 10%	0.38 ml	SDS 10 %	0.75 ml	0.75 ml

Table 3.2 Recipes for SDS-PAGE gels

Protein samples were mixed with 4x sample buffer and heated at 95°C for 3 min. In some cases urea (final conc. 6M) was added to the sample. 15 % and 18 % gels were used for effective separation range between 10-60 and 5-30 kDa, respectively. Gels were run at a constant current at 25 mA for 45-60 min and stained with Coomassie Blue for 30 min with gentle agitation. In some cases rapid silver staining was used (Nesterenko et al.,

1994). Gels were afterwards distained or used unstained for the electrotransfer in western blotting.

To estimate protein size, the Low molecular weigh-SDS Marker Kit (GE-Healthcare) was used Table 3.3:

LMW-SDS Marker Kit	MW (kDa)
α-lactalbumin	14.4
Trypsin inhibitor	20.1
Carbonic anhydrase	30
Ovalbumin	45
Albumin	66
Phosphorylase b	97

Table 3.3 The list of proteins and their sizes used for SDS-PAGE electrophoresis

3.4.8 Western blot

After the separation of samples by SDS-PAGE, western blot was done using the semi-dry-electrotransfer. At the end of electrophoresis the gel was removed and attached to a PVDF membrane that had been activated in methanol for 5 min. A 3 mm Whatman paper was soaked in blot buffer, placed on the anode and gently pressed to remove air-bubbles. Then, the activated PVDF membrane was placed on the paper. On top of the PVDF membrane the SDS-PAGE gel was placed, and another soaked 3 mm Whatman paper was placed on top. Again, air-bubbles were removed. Finally, the cathode was placed on the top and transfer was done with a current of 2-3 mA/cm^2 of gel (20 V) for 2-3 h. After the transfer the gel was stained with Coomassie Blue for 30 min. The PVDF membrane was stained with Ponceau S for 5 min. When the proteins were visible the membrane was washed with deionised water and photographed. Afterwards, the membrane was incubated for 1 h in blocking solution (M-TBS buffer). Then, membrane was incubated for 1 h with the primary antibody diluted in M-TBS buffer. After washing with M-TBS (2 x 5'), the membrane was incubated with the secondary antibody diluted in M-TBS buffer

for 1 h. Washing was repeated. For chemoluminiscence detection a mixture of 2 ml of solution A and 50 µl of solution B of the ECL-Plus western blotting detection kit was prepared. The mixture was uniformly distributed over the surface of the membrane and incubated for 5 min. Finally, a Hyperfilm-ECL chemoluminiscence-sensitive film was exposed from 5''to 5'.

3.4.9 Protein-DNA interaction assays

Agarose and polyacrylamide gels were used for detection of the differences in DNA mobility after protein binding (DNA mobility shift assay). Targeted or random DNA sequences (100-400bp) were mixed with increasing amounts of the protein sample. Mixtures were incubated for 30 minutes in binding buffer (30 mM HEPES pH 7.5, 50 mM NaCl, 1 mM DTT, 1 mM EDTA and 8-10% glycerol) and then applied on the gel. 2% agarose gel was used for a separation of DNA and the DNA-protein complex. For detection of DNA, ethidium-bromide was added in the gel to the final concentration of 0.5µg/ml,. 7-10 % polyacrylamide gels (TBE buffer) were also used to follow the migration of DNA-protein complexes. Two gels loaded with the same samples were run under same conditions in a GE-Healthcare electrophoresis chamber. TBE buffer was used as electrophoresis buffer, and was mixed between anode and catode chamber every 10 minutes to ensure constant buffer conditions (reduce the so called "smiling effect" and uneven distribution of pH through the gel.) Gels were fixated with a 0.25 % solution of glutaraldehyde. One gel was stained with ethidium-bromide (1 µg/ml) for DNA and the other with the rapid silver-staining protocol for proteins. Comparison of the two gels revealed co-migration of DNA and protein. For detection and photographing of the gels, the Biorad Chemi Doc™ gel documentation system with the Quantity One® software was used.

3.4.10 Chaperone assays

Chaperone activity of the proteins was tested in heat-induced protein aggregation assays, with firefly luciferase (Promega) and porcine citrate synthase (Sigma) as substrates. Luciferase (0.15 μM) was mixed with the respective recombinant protein in different molar ratios in assay buffer (30mM HEPES pH 7.2, 5 mM $MgCl_2$, 100mM KCl). Thermal aggregation was monitored at 45°C as an increase in attenuance at 320nm, using a PerkinElmer Lambda 25 UV/VIS Spectrometer equipped with a thermostatted cuvette holder. For the assay with citrate synthase (0.2 μM), aggregation was monitored at 50°C in assay buffer (30mM HEPES pH 7.2). In cases where AAA modules or complete AAA proteins were tested for activity, 5 mM $MgCl_2$ was added in the buffer. When indicated, ATP or other compounds (ATPγS, AMP-PNP, ADP, etc) were added to the samples. All experiments were performed three times and the curves shown in results represent the average of three measurements.

Complexes of the tested proteins and their substrates were isolated by pull-down assay using 6xhis-tags on the recombinant protein or antibodies against the substrate proteins. Luciferase or citrate synthase were incubated with the protein of interest in eppendorf tubes in the same conditions described above for 30-40 minutes. Samples were centrifuged (13000 x g, 10 minutes) and the supernatant was used for pull-down. Antibodies and resins (Ni-NTA sepharose, Protein G (A) sepharose) were added to supernatant and the mixture was incubated for 30 minutes on room temperature. Resins were washed several times with washing buffer (buffer B for his-tag, PBS-T for Protein G (A) sepharose) and the proteins were eluted with 2xSDS sample buffer. Samples were analyzed by SDS-PAGE.

3.4.11 ATPase activity assay

ATPase activity of purified AAA proteins was measured at different temperatures in 20mM MOPS/NaOH pH 7.3, 150mM NaCl, 5mM $MgCl_2$ by a modified malachite green assay for nanomolar amounts of inorganic phosphate (Lanzetta et al., 1979, Veldhoven

and Mannaerts, 1987). Protein samples were preincubated at the respective temperatures for 15 min prior to addition of ATP. Reactions were stopped by addition of EDTA (5 mM) and malachite green solution, and incubated 15 minutes at room temperature. Absorbance was measured at 650 nm. A calibration curve with KH_2PO_4 was used to determine the concentration of free phosphate produced in the assays. Values of control samples without protein were subtracted to account for spontaneous ATP hydrolysis at high temperatures.

3.4.12 Negative staining electron microscopy of the protein complexes

Macromolecules were adsorbed on the carbon-coated Pioloform support film that had been subjected to glow-discharging in a vacuum unit (Edwards Plasmaglo ion bombardement E09021000) prior to use to improve surface hydrophilicity. After adsorption the excess liquid was removed with a filter paper. The grids were washed with water to remove buffer salts and negatively stained with 1% aqueous uranyl acetate. Samples were analyzed in a Philips CM10 transmission electron microscope at 60 kV acceleration voltage using a 30 µm objective aperture at a primary magnification of 52 000x. Images were recorded on MACO EM film EMS of 4x5 inches size (Hans Mahn & Co., Stapelfeld, Germany) and scanned at 1200 dpi resolution (Epson Expression Pro 1680). Electron microscopy at low resolution was done by Heinz Schwarz at the microscopy service unit of the MPI for developmental biology, Tuebingen. High resolution microscopy and image averaging was done by Beate Rockel at MPI for Biochemistry, Dept. of Molecular Structural Biology, Martinsried.

3.5 Bioinformatics

Several servers and software were used for the bioinformatical analysis of protein sequences. BLAST and PSI-BLAST (Altschul et al., 1997) search was used to retrieve locally similar sequences with an input amino acid sequence. Search was done against protein non-redundant database at NCBI (http://www.ncbi.nlm.nih.gov/entrez). Multiple

sequence alignment of the BLAST and PSI-BLAST searches was done using MUSCLE (**mu**ltiple **s**equence **c**omparison by **l**og-**e**xpectation; Edgar, 2004), BLAMMER or MACAW (Schuler et al., 1991). Comparison of two alignments with each other was done with HHalign (http://protevo.eb.tuebingen.mpg.de/toolkit/).

 Publicly available PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/; Jones, 1999) and JNET (http://www.compbio.dundee.ac.uk/~www-jpred/; Cuff and Barton, 1999) servers were used to predict the secondary structure proteins. Quick2D, which summarizes and compares outputs of several secondary structure prediction servers, was also used (http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=psipred).

For tertiary structure prediction of proteins we used either our in-house server HHpred (http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred) or the 3-DJury structure prediction Meta server (http://bioinfo.pl/Meta/). Models of protein structures were generated automatically by servers those servers or through Swiss Model - an automated comparative protein modelling server, through an alignment interface (http://swissmodel.expasy.org/). Refinement of the modeled structure was done with software from Barry Honig's bioinformatics group (http://honiglab.cpmc.columbia.edu/). Clustering as well as graphical representation of the clustered sequences was done using the CLANS software (Frickey and Lupas, 2004). The software can be downloaded from: http://protevo.eb.tuebingen.mpg.de/cgi-bin/download/download.pl.

4. Results

<u>4.1 VAT-Nn and VAT-Nn mutants (introductory remarks)</u>

Vat-Nn represents the N-terminal subdomain of the VAT protein (gi: 6435755; amino acids 1-91) which adopts a double psi-barrel fold. A number of important questions had to be answered in connection with the evolution of this complicated fold (Appendix Fig.3). In search for the possible answers I build a series of constructs that were supposed to mimic possible intermediates that arose in near or far history of this fold (Fig. 4.1). Such intermediates would at the end help to interpret possible events that happened during the evolution.

The relationship of VAT-Nn with a postulated ancestral $\beta\alpha\beta\beta$ sequence (Appendix Fig. 2) was elucidated by constructing series of VAT-Nn half sequences with circular permutation of the first β-strand (VAT-Nnc cp) or by addition of the last strand from similar sequences like AbrB (a transcription factor from Bacillus subtilis) or Ta1217 (a hypothetical protein from *Thermoplasma acidophilum*) (VAT-Nnc+βTa1217). These constructs (Fig 4.1, violet sequences) were made to address a possible circular permutation event and to connect the fold of AbrB-like transcription factors with the family of double-psi barrels.

Since the ancestral sequence of VAT-Nn was probably created by duplication of an "ancestral half-barrel", it was important to know if the two halves of the VAT-Nn barrel (VAT-Nnc and VAT-Nnn) can fold by themselves. It was assumed, that these constructs would fold into homodimers as well as would their native counterpart, excisionase-like protein from a plasmid of *Methanothermobacter thermautotrophicus* (green group of sequences in Fig. 4.1).

The gene of the "ancestral half-barrel" would by time, in the evolution, be duplicated (VAT-Nnc/Nnn db), to result in an ancestral protein containing two "equal" halves in one polypeptide chain (Fig. 4.1, blue sequence). The fold of this ancestral protein is expected to be close to the one that is present in VAT-Nn today.

The role of the psi-loop as well as its length was examined by designing "loopless" constructs (VAT-Nn/Nnc ll or VAT-Nnc db ll – red group in Fig. 4.1). These designs

were based on sequence comparisons of the VAT-Nn sequence with presumably loopless "hypothetical transcription regulator" from *Pyrococcus h.* (Phs018). Most of the 'rational' constructs were based only on sequence similarities between VAT-Nn and other homologous proteins. Structural data was restricted, beside Vat-N (1CZ4), to the structure of AbrB-N (1EKT), which prompted us to express and structurally characterize other "native counterparts" from the sequence alignment (Fig. 1.9).

```
                                     repeat
                     |-------------------------------------------|
                         β1        Ψ       β2   α1               β3
                        sssssss          ssss hhhhh           sssss
VAT-Nn Ta            mesnnGIILRVAEANSTDPG-MSRVRLDESSRRLLDAE-IGDVVEIEKVr
                     KTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV              90
VAT-Nnc db           mvKTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKVgh
                     KTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV


                          ssss            sss hhhhhh           ssss
PhS018               mDVLAKFHTTVHR------IGRIIIPAGTRKFYGIE-QGDFVEIKIVkyegee
                     PKEGTFTARVGE------QGSVIIPKALRDVIGIK-PGEVIEVLLLghykprn      92
VAT-Nn ll            mesnnGIILRVAEAN------MSRVRLDESSRRLLDAE-IGDVVEIEKVr
                     KTVGRVYRAR------KGIVRIDSVMRNNCGAS-IGDKVKVRKV
VAT-Nnc dbll         mvKTVGRVYRAR------KGIVRIDSVMRNNCGAS-IGDKVKVRKVegeepr
                     KTVGRVYRAR------KGIVRIDSVMRNNCGAS-IGDKVKVRKV


                          sssss          sss hhhhhh         sssssssss
Mt6002757            mameDVGVPFSNRLTR----QGNIKVPADLRDALKLK-PGDLLVVEIKkvdrs       48
VAT-Nnn              mesnnGIILRVAEANSTDPG-MSRVRLDESSRRLLDAE-IGDVVEIEKVr
VAT-Nnc              mvKTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV


                          ssssss              hhhhhhh      ssssss   sssssss
Bs 1EKT/AbrB-N        MKSTGIVRKVDE-----LGRVVIPIELRRTLGIA-EKDALEIYVDD—EKIILKKYkpnmt
      53
Ta1217          mtdnkkiMDIARMTKRGA------SVRVTIPKKVLKKLNFK-DEDLIAFYESED-GRIYIDLLk    56
                                       ssss hhhhh          ssssss   sssssss
VAT-Nnc cp                     mARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV---KTVGRVYR
Vat-Nnc+βTa1217      mvKTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKVED-GRIYIDLLk
```
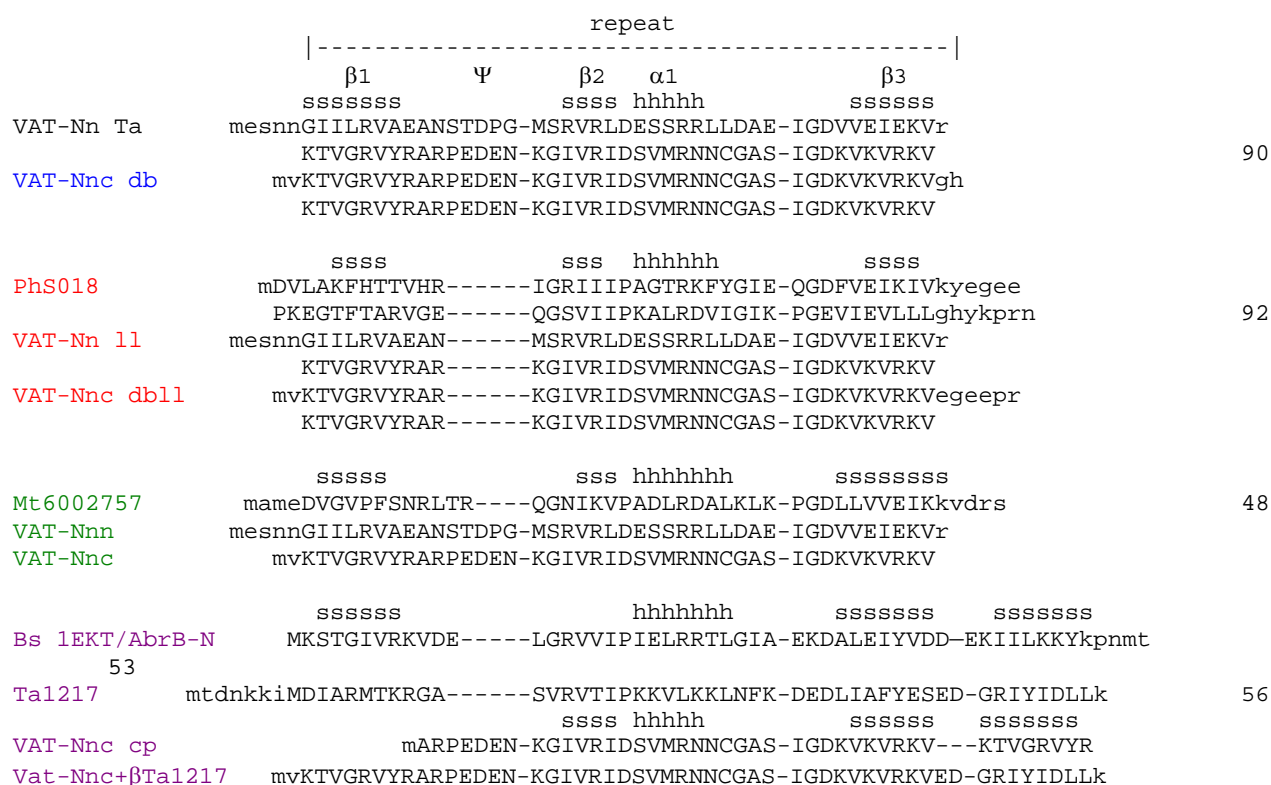
Figure 4.1 Alignment representing VAT-Nn, related sequences and constructs that were made to track possible events during the evolution of the fold as well as relations between different groups. Details of colored groups are given in text.

4.1.1 Expression and purification of VAT-Nn and halves

Expression of the VAT-Nn and VAT-Nnc constructs from the pGEX-4T1 vector resulted in very high amounts of GST-fusion proteins. The VAT-Nnn construct was expressed from the pET30b vector with a 6xHis-tag (Fig. 4.2). VAT-Nn and VAT-Nnc were purified on a GSH column and fused-GST was removed by thrombine

cleavage. Gel filtration of VAT-Nn revealed two distinguishable species that correspond by size to a monomers and dimers. VAT-Nnc showed the same patterns, presumably forming dimers and tetramers. The two species could be separated by gel filtration and behaved different in cross-linking experiment (Fig. 4.3).
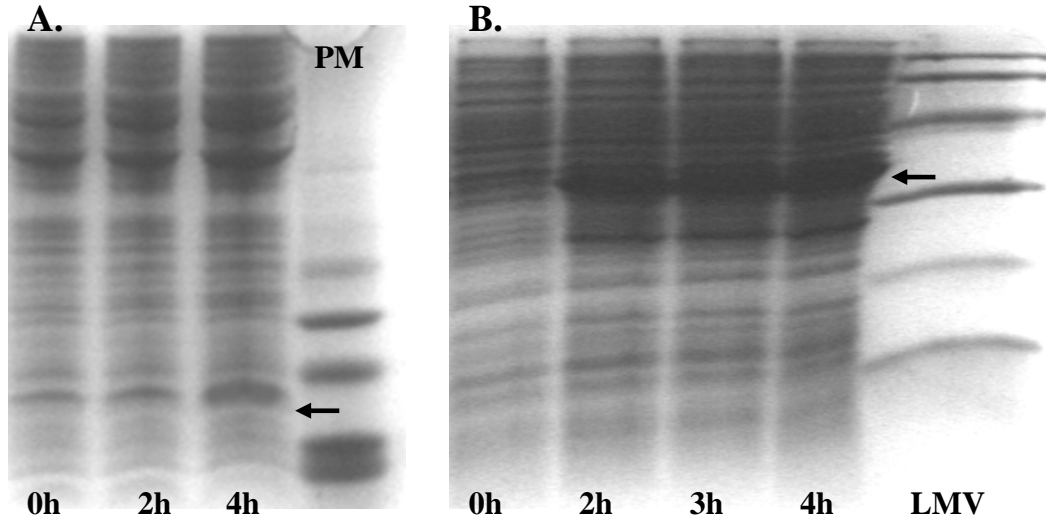

Figure 4.2 Test expressions of (**A.**) 6xHis-VAT-Nnn and (**B.**) GST-VAT-Nn proteins. Black arrows indicate expressed proteins. PM – peptide marker; LMW – low molecular weight marker
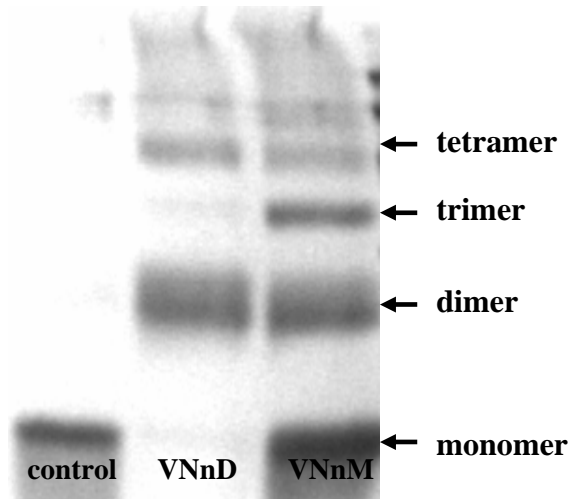
Figure 4.3 Crosslinking of VAT-Nn dimer (VNnD) and monomer (VNnM) with 0.3% glutaraldehyde. Shown on the left side is a control without crosslinker. Arrows indicate the positions of monomers, dimmers, etc.

CD spectra of purified VAT-Nn monomers and dimers and VAT-Nnc dimers and tetramers were almost identical and indicated folded proteins with both α-helix and β-strand content (Fig. 4.4).
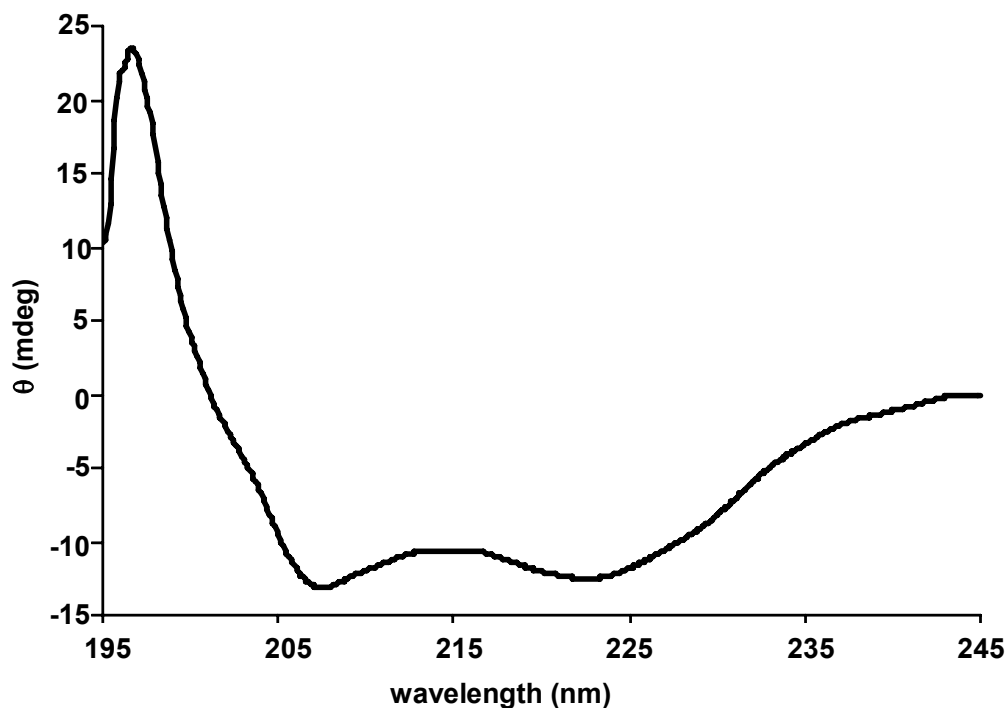
Figure 4.4 CD spectrum of VAT-Nn dimer. Typical α helical minima at 209 and 222 nm are pronounced.

VAT-Nnn was expressed and purified as a soluble protein, but its CD spectra and treatment with proteinase K indicated that the protein was not folded (data not shown).

VAT-Nnc db (blue sequence in Fig. 4.1) contained two copies of VAT-Nnc, the C-terminal half of the VAT-Nn barrel, connected by a Gly-His linker. It was expressed as a GST-fusion protein, purified on a GSH column and cleaved from the fusion protein using thrombin. VAT-Nnc db eluted as a monomer from gel sizing column and was folded according to the CD spectrum (Fig. 4.5) and resistance to proteinase K treatment. Further analysis of this protein by NMR spectroscopy revealed that it showed peaks that are characteristic for a VAT-Nn spectrum. Therefore, we assumed that it adopts the double-psi barrel fold.
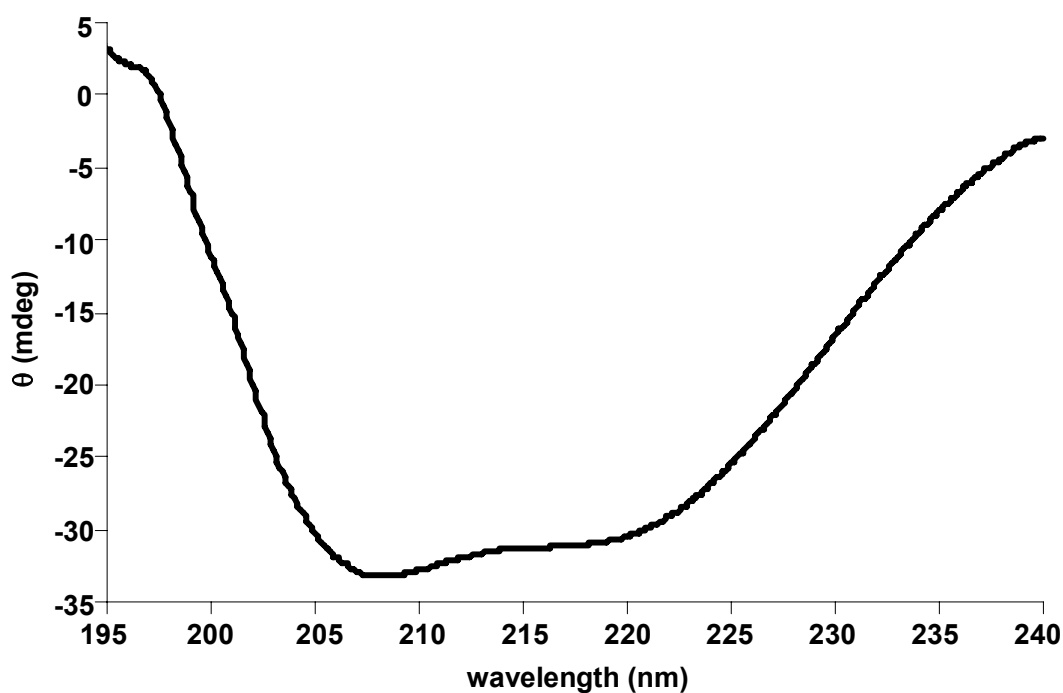
Figure 4.5 CD spectrum of VAT-Nnc duplicated (VAT-Nnc db) construct

4.1.2 Chaperone activity of VAT-Nn and VAT-Nnc

The chaperone-like activity of VAT-N (Golbik et al., 1999) prompted us to test our folded constructs (VAT-Nn, VAT-Nnc) in similar assays. VAT-N protein was expressed from the pET28b vector in C41 cells and purified over a Ni-NTA column using a C-terminal 6xHis-tag. The protein was further purified using gel filtration to achieve more than 95% purity.

Proteins of interest were mixed in different ratios with protein substrates (citrate synthase and luciferase) and heated. Aggregation was then followed by the change of absorbance at 320nm. Protein aggregation was calculated relative to the total aggregation of the protein substrate alone (Fig. 4.6).

A.

B.
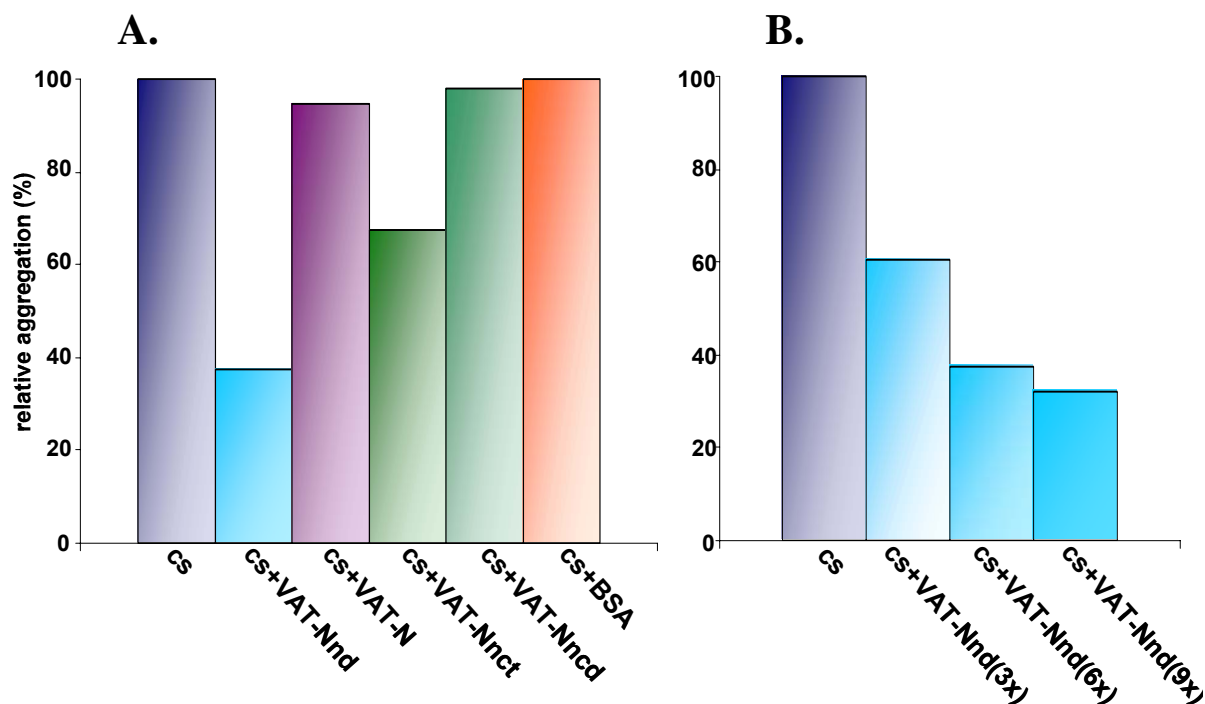
Figure 4.6(A.) Citrate synthase(cs) heat aggregation assay in the presence of different proteins. VAT-Nnd stands for the dimer of VAT-Nn; VAT-Nnc(d/t) stand for **d**imer or **t**etramer of VAT-Nnc. Assays were done with a six-fold excess of the protein of interest. Note: lower % of the relative aggregation means higher chaperone activity. **(B.)** Citrate synthase heat aggregation assay in the presence of different concentrations of VAT-Nnd. Numbers in brackets represent molar ratio of the VAT-Nnd over citrate synthase

It is noticeable that only the dimer of VAT-Nn and the tetramer of VAT-Nnc showed chaperone activity, thus preventing heat aggregation of citrate synthase (Fig. 4.6 A). It is also important that this activity could be titrated by changing the ratio of active protein versus protein substrate, indicating the requirement of the 'chaperone' protein in this process (Fig. 4.6 B). Monomers of VAT-Nn, dimers of VAT-Nnc, VAT-N and BSA, which was used as a negative control, did not show any activity.

4.1.4 VAT-Nn(c) loopless and circular permutation mutants

Several constructs were made from VAT-Nn and its halves (VAT-Nnn/VAT-Nnc) in order to reconstruct hypothetical evolutionary events and to further elucidate the importance of the psi-loop length (Fig 4.7). All constructs were expressed and purified like soluble fusion proteins with N-terminally fused GST. Upon cleavage of GST with thrombin, most of the constructs were soluble except the VAT-Nn loopless(VAT-Nn ll) construct, which precipitated immediately after cleavage and could not be refolded under any condition tested. Excision of the psi-loop either in VAT-Nnc (ll) or VAT-Nnn (ll) resulted in soluble proteins, but they were not folded based on CD spectra, proteinase K treatment and NMR spectroscopy.

```
                                           repeat
                       |-------------------------------------------|
                         β1          Ψ      β2    α1               β3
                         sssssss            ssss  hhhhh          ssssss
VAT-Nn Ta         mesnnGIILRVAEANSTDPG-MSRVRLDESSRRLLDAE-IGDVVEIEKVr
                  KTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV                  90


                         ssss                sss   hhhhh         ssss
PhS018             mDVLAKFHTTVHR------IGRIIIPAGTRKFYGIE-QGDFVEIKIVkyegee
                   PKEGTFTARVGE------QGSVIIPKALRDVIGIK-PGEVIEVLLLghykprn        92
VAT-Nn ll         mesnnGIILRVAEAN------MSRVRLDESSRRLLDAE-IGDVVEIEKVr
                  KTVGRVYRAR------KGIVRIDSVMRNNCGAS-IGDKVKVRKV
VAT-Nnc dbll        mvKTVGRVYRAR------KGIVRIDSVMRNNCGAS-IGDKVKVRKVegeepr
                  KTVGRVYRAR------KGIVRIDSVMRNNCGAS-IGDKVKVRKV
VAT-Nnc ll         mvKTVGRVYRAR------KGIVRIDSVMRNNCGAS-IGDKVKVRKV


Ta1217       mtdnkkiMDIARMTKRGA------SVRVTIPKKVLKKLNFK-DEDLIAFYESED-GRIYIDLLk    56

                                     ssss hhhhh         ssssss  sssssss
VAT-Nnc cp              mARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV---KTVGRVYR
Vat-Nnc+β          mvKTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKVED-KTVGRVYR
Vat-Nnc+βTa1217    mvKTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKVED-GRIYIDLLk
VAT-Nncll+βTa1217  mvKTVGRVYRAR------KGIVRIDSVMRNNCGAS-IGDKVKVRKVED-GRIYIDLLk
```

Figure 4.7 Alignment of natural sequences of VAT-Nn, Phs018 and Ta1217, with loopless (red) and circular permutated (violet) constructs made from VAT-Nn(c) to resemble natural sequences and to elucidate the importance of differences on the sequence levels.

Duplication of the VAT-Nnc loopless construct with "the connector" (EGEEP) from the Phs018 protein resulted in soluble and folded protein as judged by CD (Fig. 4.8)

and NMR spectra. Structure determination of this protein by NMR spectroscopy is an on-going project.

Constructs that had circular permutations, addition of the first β-strand to the C-terminus (VAT-Nnc cp and VAT-Nnc+β), or addition of the last β-strand from Ta1217 (VAT-Nnc+β-Ta1217 and VAT-Nnc ll+β-Ta1217) were soluble after cleavage from the GST-fusion with thrombin. But CD and NMR spectra indicated that these proteins were not folded (data not shown).
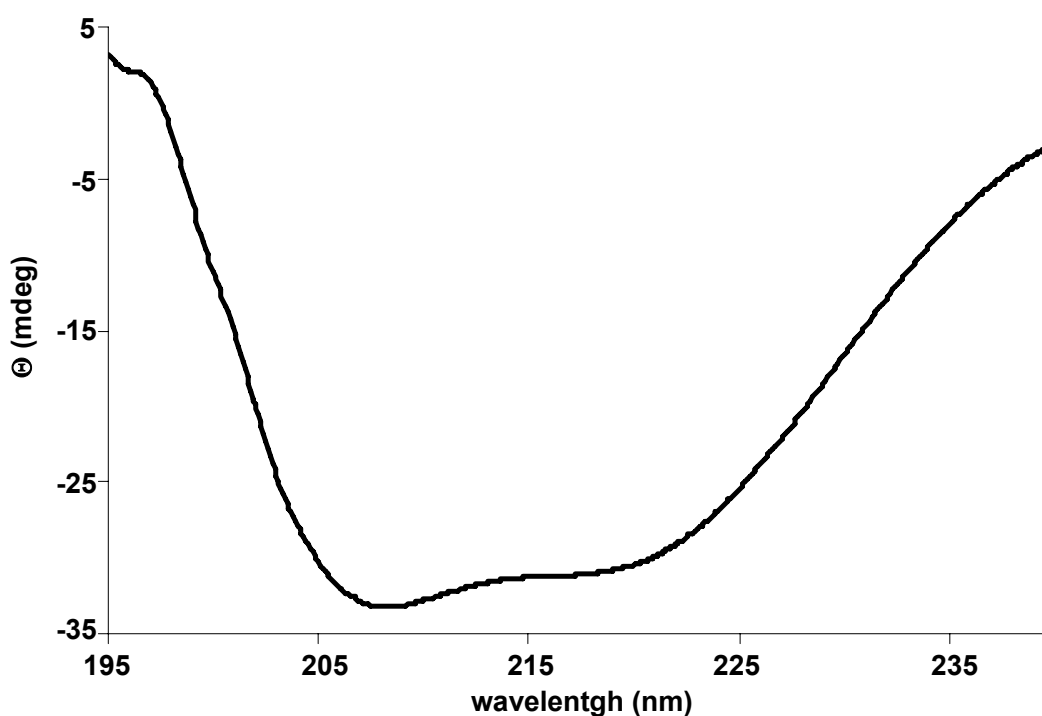

Figure 4.8 CD spectrum of VAT-Nnc duplicated loopless protein

4.1.5 Expression and characterization of Ph1179 and MT6002752. DNA binding activity

Ph1179 (gi 14591004) of *Pyrococcus horikoshii* was generated synthetically from primers by PCR and was cloned into the pET30b vector. Expression in *E. coli* at 37°C yielded insoluble protein that was purified under denaturing conditions in 6 M urea on a SP Sepharose FastFlow (Fig 4.9) cation exchange chromatography column (GE Healthcare) and refolded by dialysis against 50 mM Tris, pH 7.4 with 200 mM NaCl at ambient temperature.
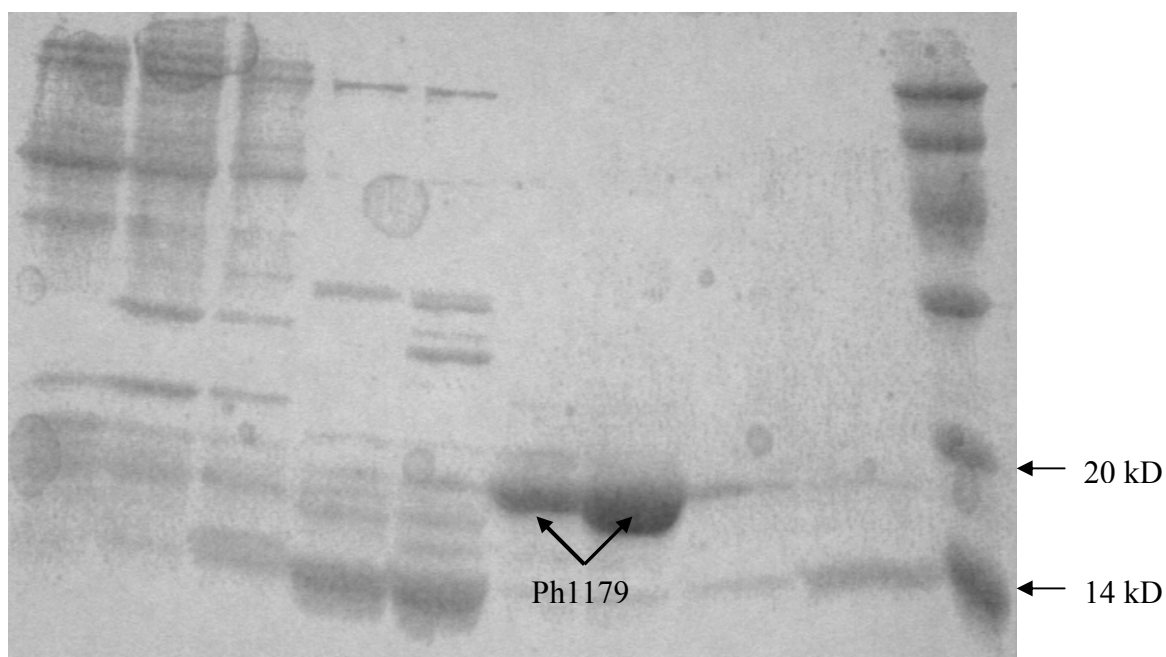
Figure 4.9 Purification of the Ph1179 protein under denaturing conditions on SP FastFlow cation exchange (GE Healthcare). Fractions were protein was eluted are with 0.7-0.8 M NaCl.

Mt6002757 (gi 6002752) of *Methanothermobacter thermautotrophicus* was also generated synthetically from primers by PCR and was cloned into the pET30b vector with an N-terminal His$_6$-tag. Additionally, a construct was made by PCR that contained duplicated Mt6002757 sequence connected by a linker encoding for Gly-His (Mt6002757db). Expression in *E. coli* at 37°C yielded in both cases insoluble protein that was purified under denaturing conditions in 6 M guanidinium hydrochloride on a Ni-NTA Sepharose chromatography column (GE Healthcare) and refolded by dialysis against 30 mM sodium phosphate buffer, pH 7.4 with 250 mM NaCl in the presence of heterologous DNA at ambient temperature. CD spectra of the refolded proteins were recorded, which indicated folded proteins with a mixture of α-helices and β-strands (Fig. 4.10).
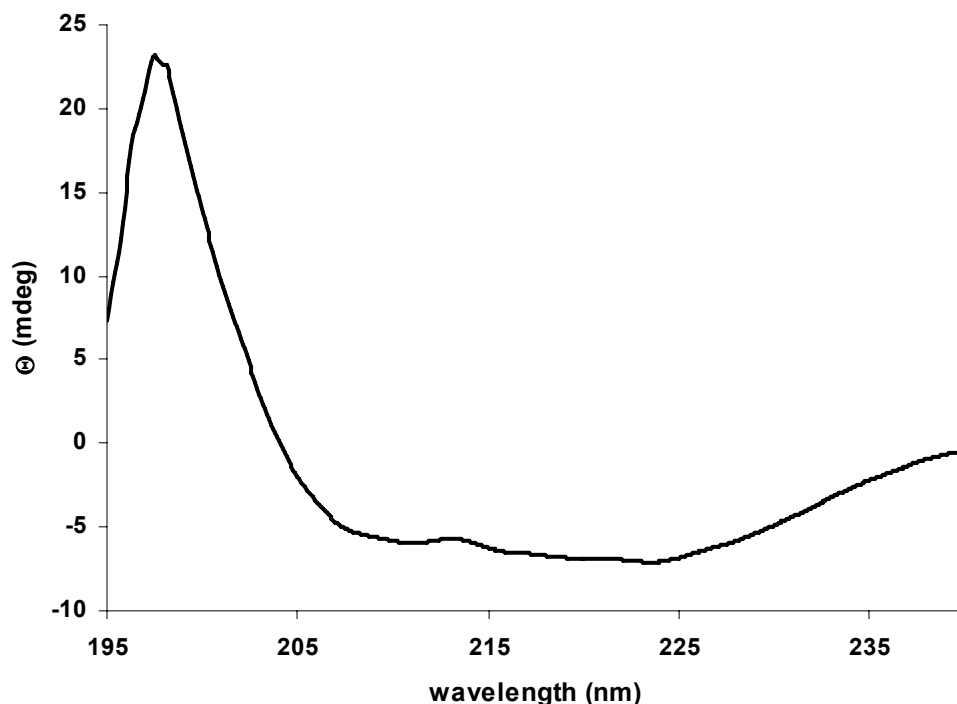
Figure 4.10 CD spectrum of Mt6002757db protein after refolding.

The purified proteins Ph1179 and Mt6002757 were analyzed on a calibrated Superdex G-75 size exclusion chromatography column (GE Healthcare) for their oligomeric state in 30 mM sodium phosphate buffer pH 7.4 and 250 mM NaCl. Ph1179 migrated like a dimer, but had the tendency to further oligomerize at higher concentrations, which made this protein not suitable for NMR spectroscopy. Nevertheless we recorded several ^{15}N spectra (Fig. 4.11), which showed characteristic peaks for VAT-Nn spectra. This prompted us to solve the structure of the homologous protein Phs018 (Coles et al., in preparation, 2GLW) adopting a RIFT barrel fold, which is related to the double psi-barrel. RIFT barrel is named for its occurrence in **Ri**boflavin synthases, **F**1ATPase and **T**ranslation factors. Based on the calibration of the gel filtration column, Mt6002757 was migrating like a dimer, while the doubled construct had the size of a monomer, showing no tendency for further oligomerization. Both proteins were too unstable for structure determination, but still enough data could be collected by NMR spectroscopy to assume that the fold is similar to Phs018.
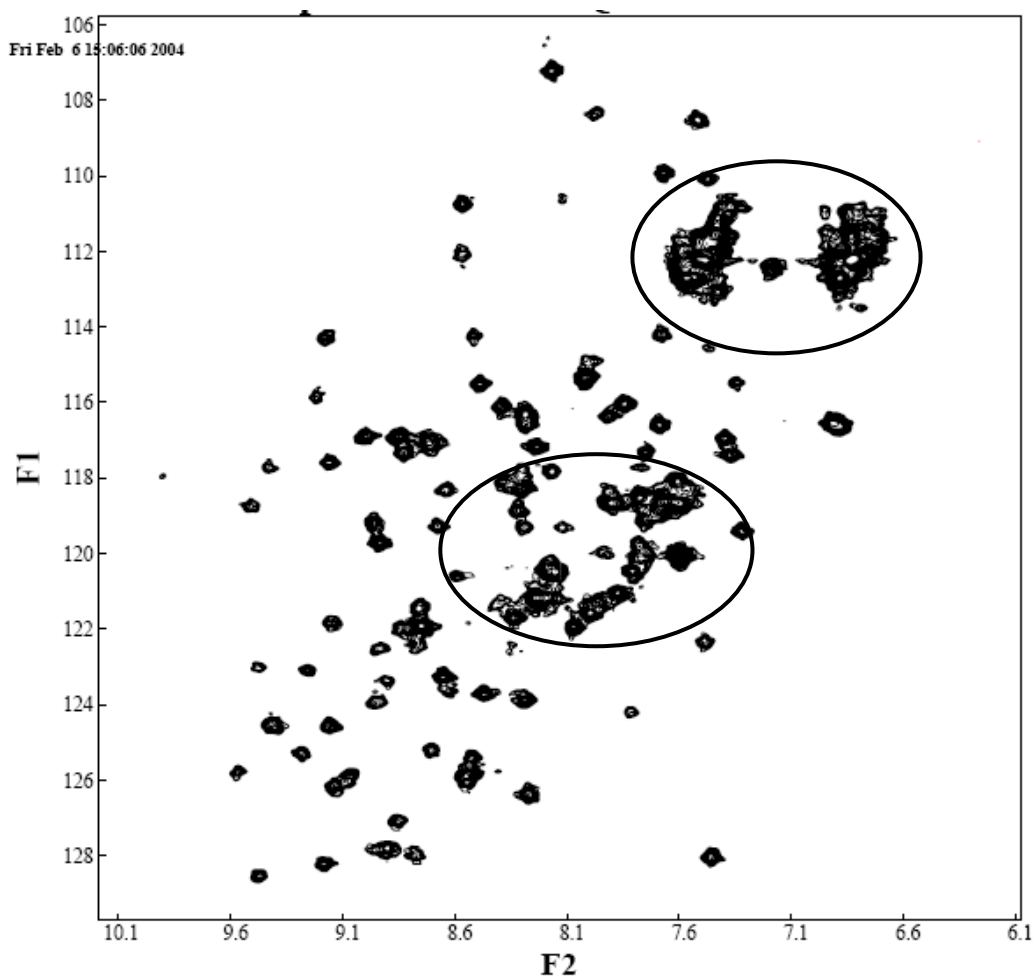
Figure 4.11 ^{15}N HSQC spectra of Ph1179. Circles show signals with low resolution that were the product of protein aggregation

Because of their sequence homology with the *Bacillus subtilis* transcription factor AbrB, unspecific DNA binding of the proteins was tested. Both Ph1179 and Mt6002757 had DNA binding activity. The VAT-Nn monomer was meant to function as a negative control, because it has protein-binding and chaperone activity. Surprisingly, the VAT-Nn monomer did also bind DNA in an unspecific manner (Fig. 4.12), suggesting an intrinsic DNA binding activity of the double-psi barrel fold. This observation was further explored in titration experiments, which gave similar results (data not shown).
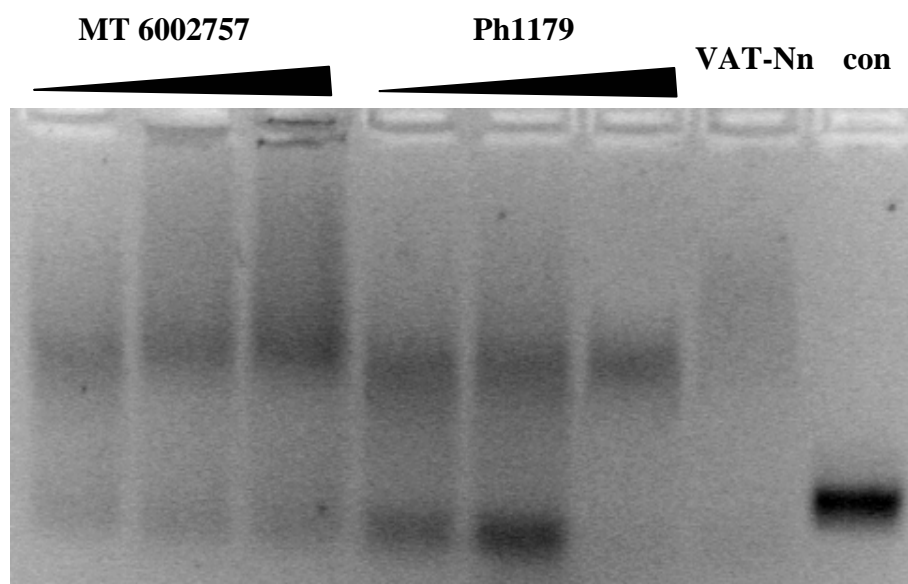
Figure 4.12 DNA shift mobility assay in presence of increasing amounts of the Mt6002757 and Ph1179. Monomeric form of VAT-Nn surprisingly also binds DNA. Control (con) was DNA fragment in binding buffer without addition of protein.

Bioinformatic analysis using PSI-BLAST (Altschul et al., 1997) in order to find homologous sequences to Mt6002757 did not give any results. Therefore, this protein was annotated as a singleton. Yet, recently we found several environmental sequences that share some sequence homology with the Mt6002757 protein.

A bioinformatic analysis of Ph1179 using the gene annotation and analysis tool "The SEED" (http://theseed.uchicago.edu) revealed homologous sequences. Besides the gene for Phs018 from *Pyrococcus horikoshii* (Coles et al., unpublished), there were six genes in *Thermococcus kodakaraensis*. Among these, four were in one single gene locus (Fig. 4.13). One sequence was found in the *Pyrococcus abysii* genome. Gene loci in the different genomes did not have high similarity, except for three hypothetical proteins (numbered 2, 3 and 4 in Fig. 4.13) with no specified function and rather low sequence complexity.
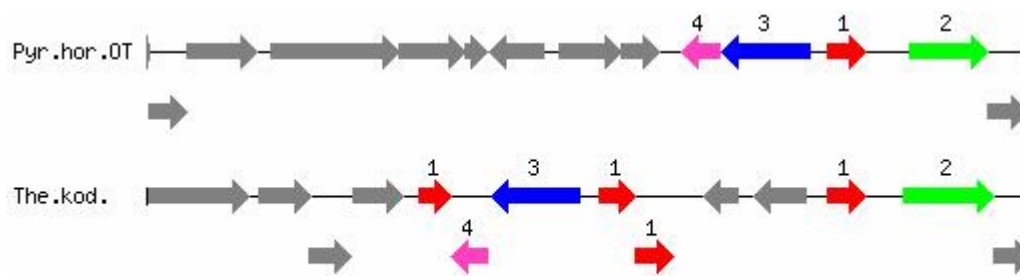
Figure 4.13 Output of a search for homologous proteins with "the SEED", using Ph1179 as query. Ph1179-like genes are labeled with number 1 and colored red. Genes labeled with different colors and numbers are conserved in the gene loci with Ph1179-like genes. (Pyr.hor.OT - *Pyrococcus horikoshii* OT3; The.kod. - *Thermococcus kodakaraensis*).

4.1.6 Expression and purification of the AbrB-N homolog –Ta1217

Ta1217 (gi16082225) of *Thermoplasma acidophilum* was generated synthetically from primers by PCR and was cloned into the pET30b vector. A 6xHis-tag was introduced at the N-terminus. Expression in *E. coli* at 37°C yielded soluble protein that was purified on a Ni-NTA column (GE healthcare), pooled and dialyzed against 50 mM TRIS/HCl, pH 7.4 with 250 mM NaCl at ambient temperature. Cross-linking using 0.15 % glutaraldehyde after 5 or 10 minutes gave 90% of the protein in form of a dimer, which was also observed by gel filtration on a calibrated gel-sizing column. A DNA mobility shift assay under non-specific conditions and in the presence of increasing amounts of Ta1217 clearly showed an interaction of the protein with DNA (Fig. 4.14).

After storage at 4°C the majority of the protein was precipitated and could be resolubilize in 6M guanidinum. Refolding was done by overnight dialysis against 50 mM TRIS/HCl, pH 7.4 with 250 mM NaCl buffer at ambient temperature. Cold sensitivity as well as the fact that it could not be concentrated to higher concentrations (max. 0.05 mM) made this protein not suitable for structure determination by NMR spectroscopy.
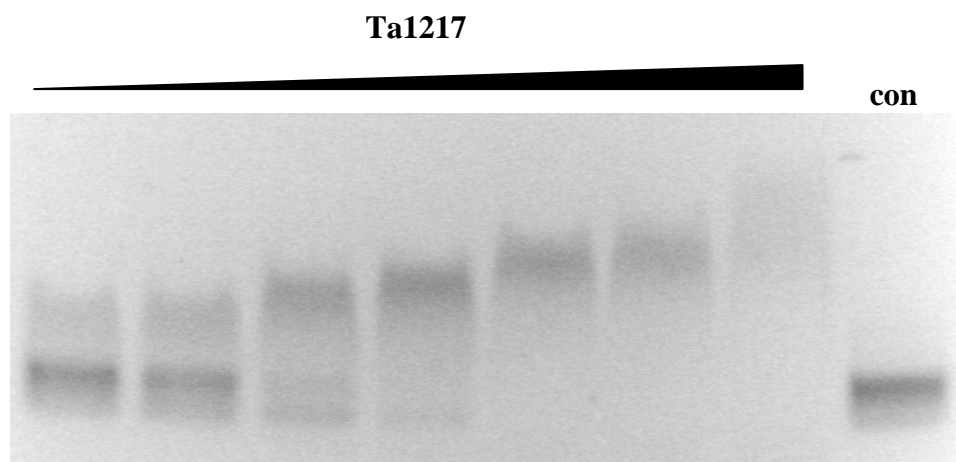
Ta1217

con

Figure 4.14 DNA mobility shift assay in the presence of different concentrations of the Ta1217 protein. On the far right side is a DNA control (con) without protein.

4.1.7 NMR structure of AbrB-N

Since we could not determine the structure of Ta1217, we decided to re-determine the structure of AbrB-N, the N-terminal domain of a transcription factor from *Bacillus Subtilis*. The structure was published to adopt a looped-hinge helix fold, which we considered wrong because of the homology of AbrB-N to proteins like MazE or MraZ, which adopt a fold resembling the double-psi barrels (Coles et al., 2005). The AbrB-N construct (encoding amino residues 1-53 of AbrB, gi113009) was amplified from *B. subtilis* PY79 chromosomal DNA by polymerase chain reaction (PCR) and cloned into the pet30b vector (Novagen). The construct contained a 6xHis-tag at the amino terminus to facilitate purification. For expression in *E. coli* C41 strain, cells were grown in LB medium at 37°C, induced at $OD_{600}\sim0.6$ with 1 mM IPTG and harvested after 4h. Uniformly ^{15}N- or ^{13}C-labeled AbrB-N was made by growing bacteria in M9 minimal medium, using $^{15}NH_4Cl$ (0.7 g/l) and $^{13}C_6$-glucose (2 g/l) as sole nitrogen or carbon sources. A mixed ^{15}N- and ^{13}C-labeled AbrB-N was prepared by combining equal amounts of harvested cells prior to lysis. Proteins were purified by a combination of immobilized metal affinity chromatography (IMAC), ion exchange and gel sizing chromatography. For NMR measurements, samples were

concentrated to 8 mg/ml in buffer containing 20 mM potassium phosphate, 50 mM KCl, 0.02% (w/v) NaN$_3$, pH 5.8. The sample purified natively from mixed cell pellets behaved the same in NMR like a sample that was mixed, purified under denaturing conditions in 8M urea and then refolded, indicating an equilibrium in the folding-unfolding process of the AbrB-N dimer. The structure of AbrB-N (1YFB and 1YSF) was solved by Murray Coles and Vincent Truffault with an RMSD for the final set of 20 dimeric structures (residues G7-Y50) of 0.27 Å for backbone atoms and 0.71 Å for all heavy atoms (Fig. 4.15 and 4.16).
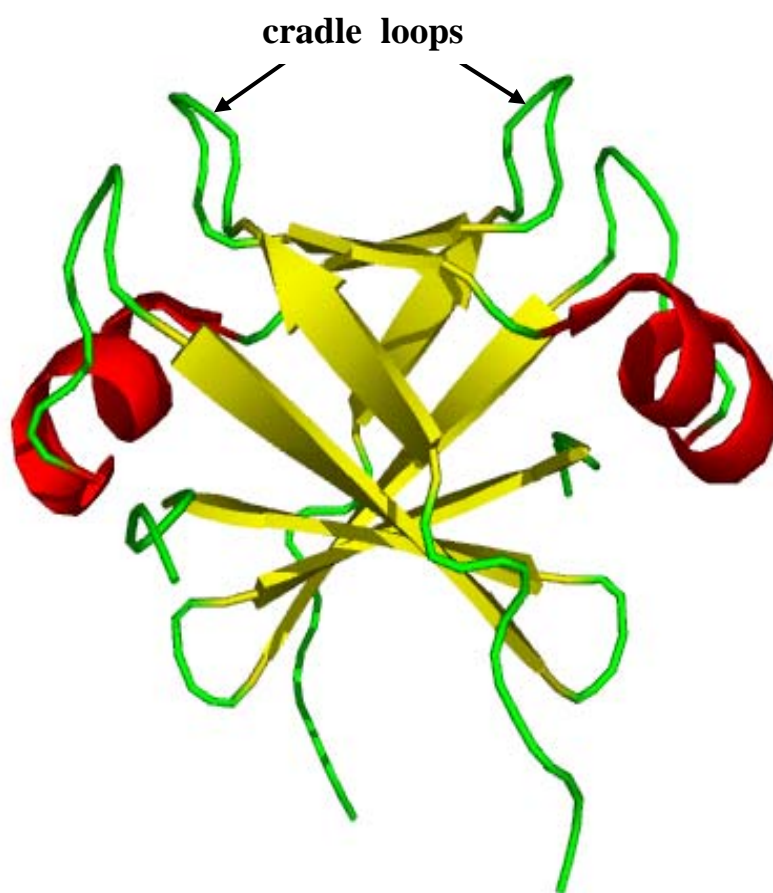
Figure 4.15 Structure of the AbrB-N in side view (1YFB). Black arrows indicate position of the cradle loops analogous to the psi-loops of VAT-Nn
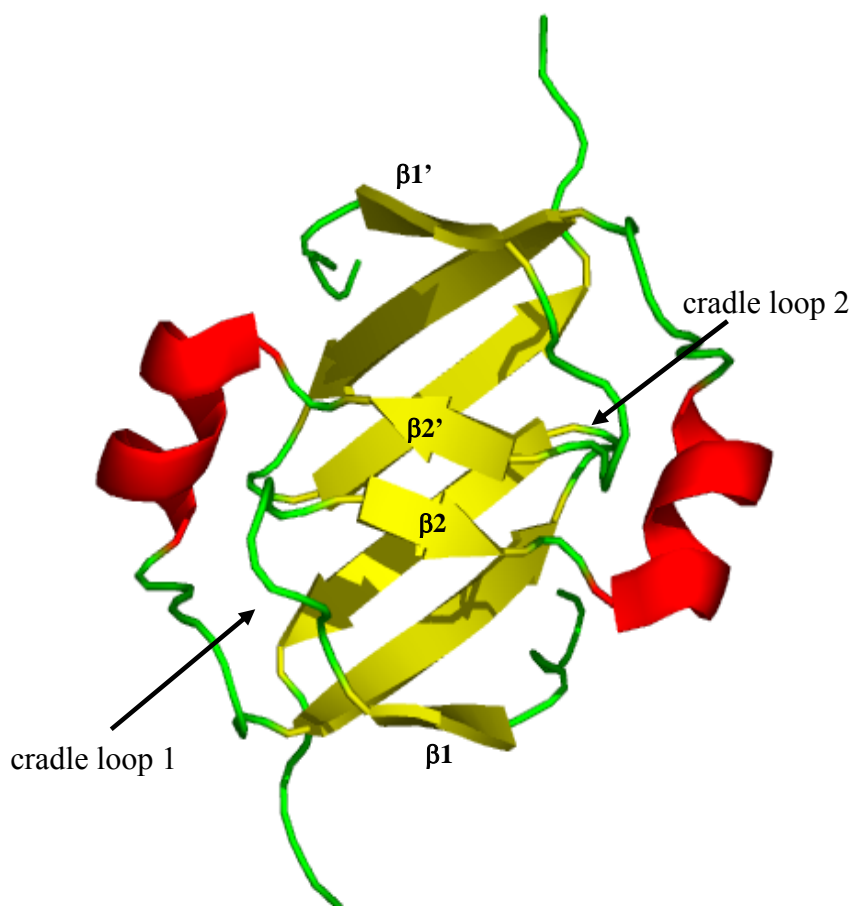
Figure 4.16 Structure of the AbrB-N in top view (1YFB). Strands flanking the cradle-loops loops are labeled β1 and β2, for the first chain, and β1' and β2' for the second. Loops are indicated by arrows.

The protein dimerizes by interleaving the four β-hairpin elements, such that each one only makes contacts to those of the dimeric partner. The result is an eight-stranded, swapped-hairpin β-barrel closed at each end by a helix (Fig. 4.16). It is noticeable that the loops between β1 and β2 (β1' and β2') project above the surface of the barrel, giving the protein a characteristic horned profile (Fig. 4.15). Overall similarity between the AbrB-N structure and the double-psi barrel fold is striking, even though these are classified into different folds. The fold presented here is markedly different from the one originally presented by Vaughn et al., where β2 and therefore one β hairpin was absent and no interleaving of monomeric elements occurred (Fig. 1.10A). The structure of AbrB-N rather resembles the fold of MazE (Fig. 1.10B), as was predicted before by bioinformatics.

4.1.8 Bioinformatic analysis of AbrB homologues

Search for sequences homologues to AbrB was done using HHsenser, a method that combines the PSI-BLAST search strategy SENSER (Koretke et al., 2002) with a method for comparing profile Hidden Markov Models (HHsearch; Soeding, 2004). After convergence of the sequence searches, which resulted in 724 sequences (580 unique proteins), the obtained sequences were clustered using the program CLANS (Frickey & Lupas, 2004) at a P-value cutoff of e-4 (Fig. 4.17).
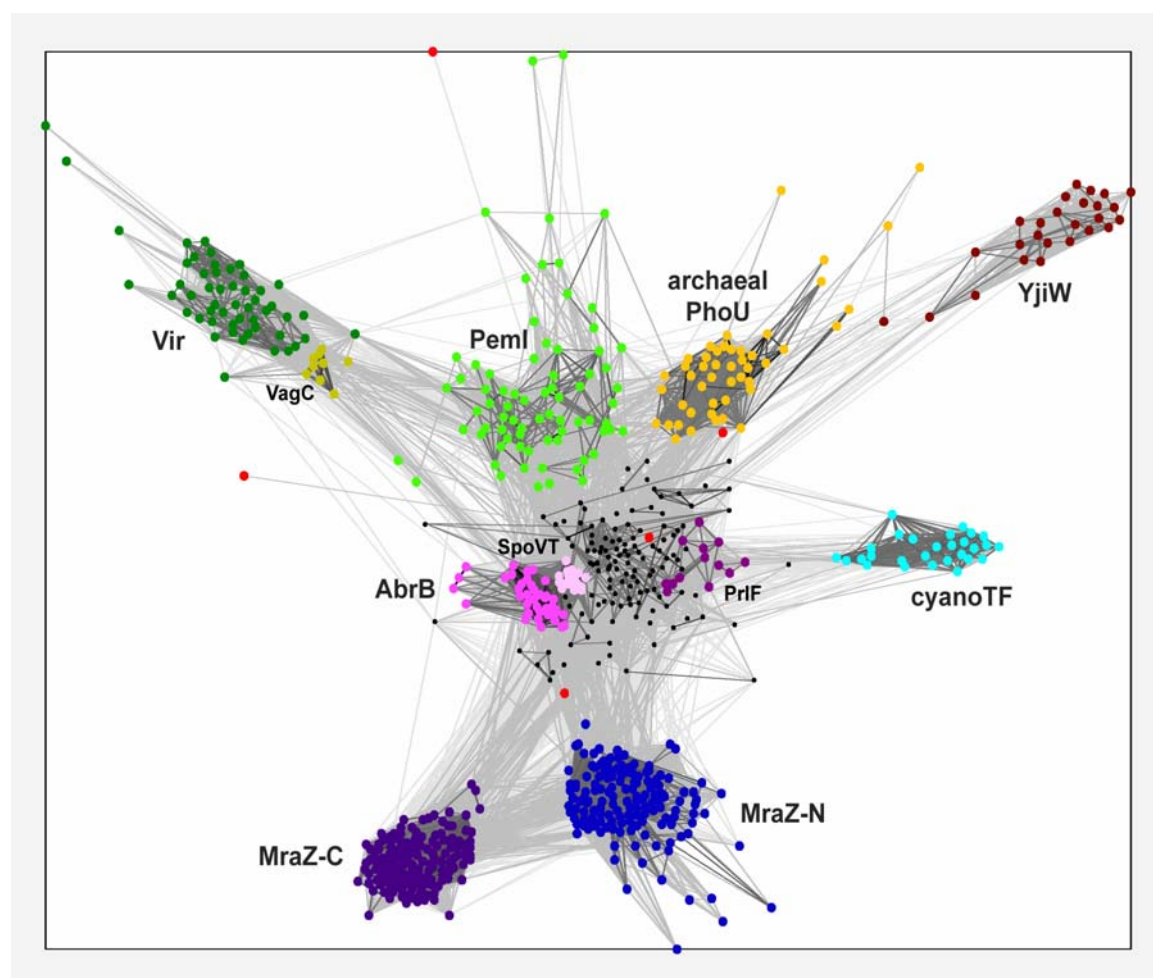

Figure 4.17 Cluster analysis of the AbrB superfamily. The eight main clusters are highlighted in different colors. Subclusters of the core AbrB cluster (black) and Vir (green) are labeled differently (AbrB:AbrB, SpoVT and PrlF; Vir: Vir and VagC). The different groups and singleton sequences (red) are discussed later.

The clustering is based on those parts of the proteins that are homologous to AbrB, which were almost always fragments of the complete sequences. In some proteins like MraZ, two fragments homologous to AbrB were recovered and named MraZ-C and MraZ-N, where C and N stands for the fragment close to either carboxy or amino terminus. The individual clusters were reexamined manually, as well as all proteins that were not clearly assigned to one cluster, but no false positives were detected. More sequences could be identified by searching with each cluster as the starting point. Some of the outliers in the clusters were sequences with a misassigned start codon in the database. These shortened AbrB domains had lower P-values. This problem is seen particularly clearly in the MraZ clusters, where the N-domain cluster is much more irregular than the C-domain cluster.

4.1.9 Bioinformatic analysis of the SpoVT sequence

SpoVT (Stage V sporulation protein T) is a transcription factor involved in positive and negative regulation of sigma-G-dependent genes during the process of sporulation in *Bacillus subtilis*. PSI-BLAST analysis using the SpoVT sequence (gi 586883) converged with 22 unique sequences – all in sporulating bacteria (see Appendix Fig. 4). Analysis of the gene locus using the already mentioned gene annotation/analysis server "The SEED" showed clear conservation through *Bacilli* and related species (Fig. 4.18). Additional sporulation genes were found in the same locus (numbered 2, 6, 7, 13 and 15 in Fig. 4.18), as well as genes for conserved proteins with different cellular functions (3, 4, 5, 9, 10, 11, 14). The gene (8) encoding the MazG protein, a nucleoside triphosphate pyrophosphohydrolase, was constantly found in the neighborhood.

Interestingly, the sequence identity in the first 55 residues of SpoVT proteins from different bacteria is in the range of 95%, while it drops to 25-30 % in the C-terminal-125 residue domain (see Appendix Fig 3.).

Figure 4.18 Analysis of the SpoVT gene locus using "The SEED" server. SpoVT genes are labeled with red color and number 1. Analysis involved several genomes of sporulating gram-positive *Bacilli* species: *Bacillus* (*subtilis*, *anthracis* and *licheniformis*), *Geobacillus* (*caustophilus* and *stearotermophilus*) and *Oceanobacillus iheyensis*.

Further analysis of the protein sequence was done using structure prediction servers at BioInfoBank Institute-MetaServer (http://bioinfo.pl/meta/) or our in-house HHpred server (http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred). The first 51 residues of SpoVT share 60% sequence identity with the AbrB-N sequence (Fig. 4.19), suggesting that it also adopts a cradle-loop barrel fold.

```
Q SpoVT      1 MKATGIVRRIDDLGRVVIPKEIRRTLRIREGDPLEIFVDRDGEVILKKYSP   51 (178)
Q Consensus  1 MKstGivRKiD~LGRVVIPkEiRrtL~I~e~D~LeifVd~d~~II1kKY~P   51 (178)
               ||||.||||||.||+||.|||||.|||+||+.|+|+++|+||||||+|.+||++||+|
T Consensus  3 MKstGivRKiD~LGRIVIPkElRrtL~I~~kd~lei~Vd~~~~IIL~KY~p   52 (96)
T AbrB        3 MKSTGIVRKVDELGRVVIPIELRRTLGIAEKDALEIYVDDE-KIILKKYKP   52 (96)
```

Figure 4.19 Alignment of SpoVT N-terminal part (1-51 aa) with the sequence of AbrB-N. Red and yellow colored boxes represent parts of the sequence that adopt β-strands or α-helices, respectively.

The C-terminal part of SpoVT (residues 52-178) is predicted with high significance by both servers to adopt the GAF or the very similar PAS domain fold (Fig. 4.20). The name GAF comes from "**c**GMP-specific and –stimulated phosphodiesterases, *Anabaena* **A**denylate cyclases and *Escherichia colt* **F**hlA"(Aravind and Ponting, 1997). The PAS domain was named after three proteins it occurs in: **p**eriod circadian protein (Per)*,* **A**h receptor nuclear translocator protein (Arnt) and **s**ingle-minded protein (Sim). PAS and GAF domains are involved in many signaling pathways where they function as sensor domains for small signaling molecules like cGMP.

```
Q SpoVT         52  ISELGDFAKEYADALYDSLGHSVLICDRDVYIAVSGSSKKDYLNKSIS-EMLERTMDQRSSVLESDAKSVQLVNGIDEDM  130 (178)
Q Consensus     52  I~EL~dFakeYAeal~~s~gh~v~I~DrD~~IaVaG~~KKey~~k~is-~~~e~~me~rks~~~~~~~~~~~~~~~~e~~  130 (178)
                    +..|..-+....+++.+...-.+.+-|.+.++..+....-.-..-+++ ..+..+.+++++.+..+...-.--...+...
T Consensus     45  ~~~L~~~~~ll~~~l~~~d~~giyl~d~~~l~l~a~~g~~~~~~ip~G~Gi~G~v~~~g~~iiv~Dv~~~~~~~~~~~~~  124 (195)
T GAF/PAS domain 45  LATLANTSALLYERLTDINWAGFYLLEDDTLVLGPFQGKIACVRIPVGRGVCGTAVARNQVQRIEDVHVFDGHIACDAAS  124 (195)


Q SpoVT        131  NSYTVGPIVANGDPIGAVVIFSKDQ-TMGEVEHKAVETAAGFLA  173 (178)
Q Consensus    131  ~~y~i~PIia~GDpIGaVii~sk~~-~~gevE~K~aetAa~FL~  173 (178)
                    .|.,..||+.++..||...+.|.+. .+.+-++.+.++-|.-++
T Consensus    125  ~S~l~vPL~~~~~viGVL~v~s~~~~~FteeD~~lL~~lA~qiA  168 (195)
T GAF/PAS domain 125 NSEIVLPLVVKNQIIGVLDIDSTVFGRFTDEDEQGLRQLVAQLE  168 (195)
```

Figure 4.20 Alignment of the SpoVT C-terminal domain (52-173aa) with the sequence of the GAF domain from the *E. coli* YebR protein (pdb code: 1VHM). Red and yellow colored boxes represent part of the sequence that adopt β-strand or α-helix, respectively.

Based on sequence similarities with the YebR protein from *E. coli* (1VHM), a model of the SpoVT GAF domain was constructed using the alignment interface of the SWISS MODEL server (http://swissmodel.expasy.org/). The structure of the model was refined using the software "Loopy" from Barry Honig's bioinformatics group at Columbia University (http://honiglab.cpmc.columbia.edu/). The software created 20 different models based on loop modeling and energy minimization of the complete structure. Models were ranked from 1-20 based on the computed energies and the best model is shown in Figure 4.21.

Figure 4.21 Model of the predicted SpoVT GAF domain based on the alignment to the YebR protein (1VHM) and refined with the software "Loopy".

4.1.10 Crystal structure of SpoVT

SpoVT and SpoVT-C (residues 57-178) were amplified from *B. subtilis* PY79 chromosomal DNA by polymerase chain reaction (PCR) and cloned into the pET30b vector (Novagen). Both constructs contained an N-terminal 6xHis tag to facilitate purification. For expression in the *E. coli* C41 strain, cells were grown in LB medium at 37°C, induced at OD_{600}~0.6 with 1 mM IPTG and harvested after 4h. Proteins were purified by a combination of Ni-NTA affinity chromatography and gel filtration on the Sephadex G-75 (GE Healthcare) column. Purity of the proteins (>95%) was checked using SDS-PAGE. For crystallization the buffer was changed by dialysis to 10 mM MOPS, pH 7.2, 100mM NaCl. Crystallization was done by the hanging drop method using a 1:1 volume ratio of protein solution to the precipitant. Plates and crystal screens from Hampton Research were used. Crystals were obtained in several conditions with both the full length protein and the SpoVT-C domain (Fig. 4.21).

A.
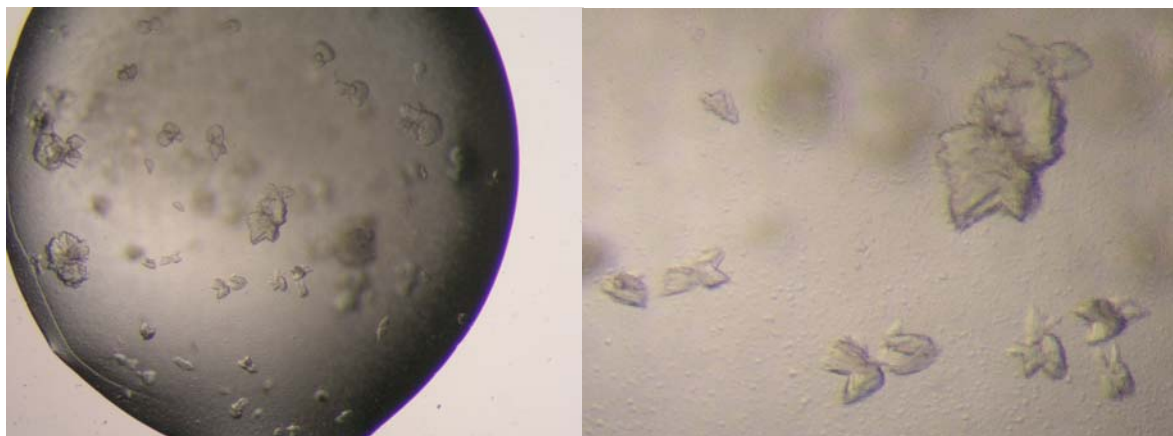
B.
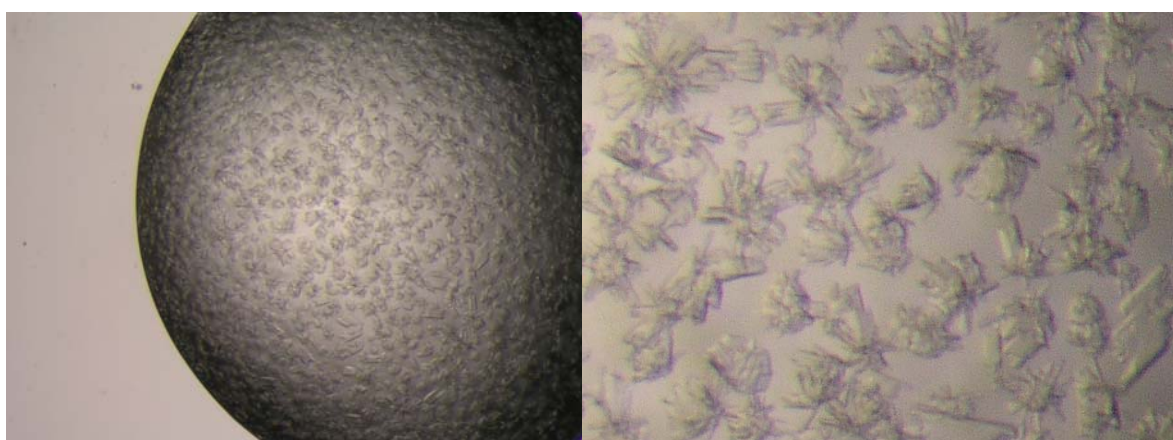
Figure 4.21(A.) Crystals of the SpoVT protein (c = 28mg/ml, 2.0 M $(NH_4)_2SO_4$, precipitant 5% v/v iso-propanol); **(B.)** Crystals of the C-terminal domain of SpoVT (c = 20mg/ml, 0.1 M HEPES, pH 7.5, 0.2 M NaCl, precipitant 25 % w/v PEG 3350)

The structure of the SpoVT C-domain (residues 57-178) was determined at 1.8 Å resolution by Iris Asen (MPI for Biochemistry, Martinsried) from crystals of the Se-Met derivate grown using 25 % w/v PEG 3350 as precipitant, 0.1 M HEPES, pH 7.5 buffer and 0.2 M $MgCl_2$. The SpoVT-C domain crystallized as a dimer, as had been observed before in size exclusion chromatography. Two GAF domains dimerized via a four-helix bundle (Fig. 4.22) built from the first (α1, α1') and last (α4, α4') helix of each chain.
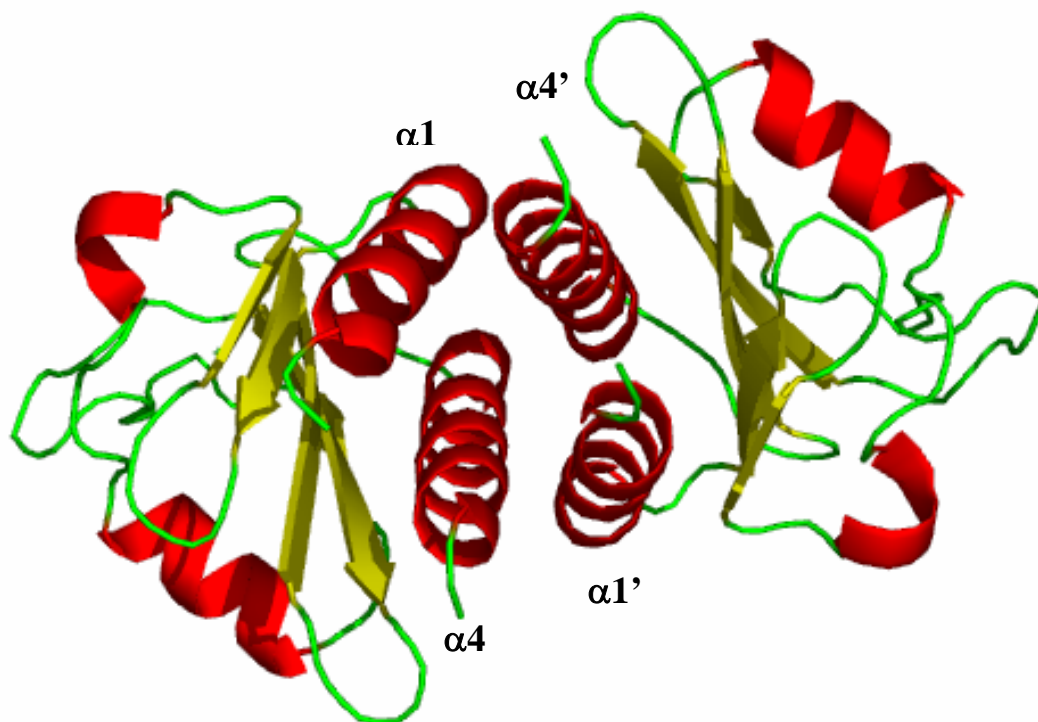
Figure 4.22 Structure of the dimeric GAF domain of SpoVT. Helices forming the four-helix bundle between two chains are labeled (α1, α4 and α1', α4' for chain A and B, respectively)

The model that was created by bioinformatics tools had a root-mean-square deviation (RMSD) from the actual crystal structure of 1.8 Å calculated on the overall structure. Models that were generated directly by structure prediction servers (HHpred, MetaServer) and without additional loop modeling and energy minimization had higher RMSD (>3.5 Å).

The crystal structure of the complete protein was solved by Iris Asen (MPI for Biochemistry, Martinsried) from a dataset of 2.5 Å resolution, using an average NMR structure of AbrB-N (1YFB) for molecular replacement of the N-terminal domain (Fig. 4.23). The structure of the full protein reveals a tetrameric protein complex via the mentioned four helix bundle formed by C-terminal GAF domains, and additionally via their N-terminal domains, which form an AbrB-N-like cradle-loop barrel (Fig. 4.23A). The GAF domains are tilted approximately 45° in respect to each other. This tilting angle is further transferred to the orientation of the cradle-loops in the swapped-hairpin barrel domains (Fig. 4.23B).
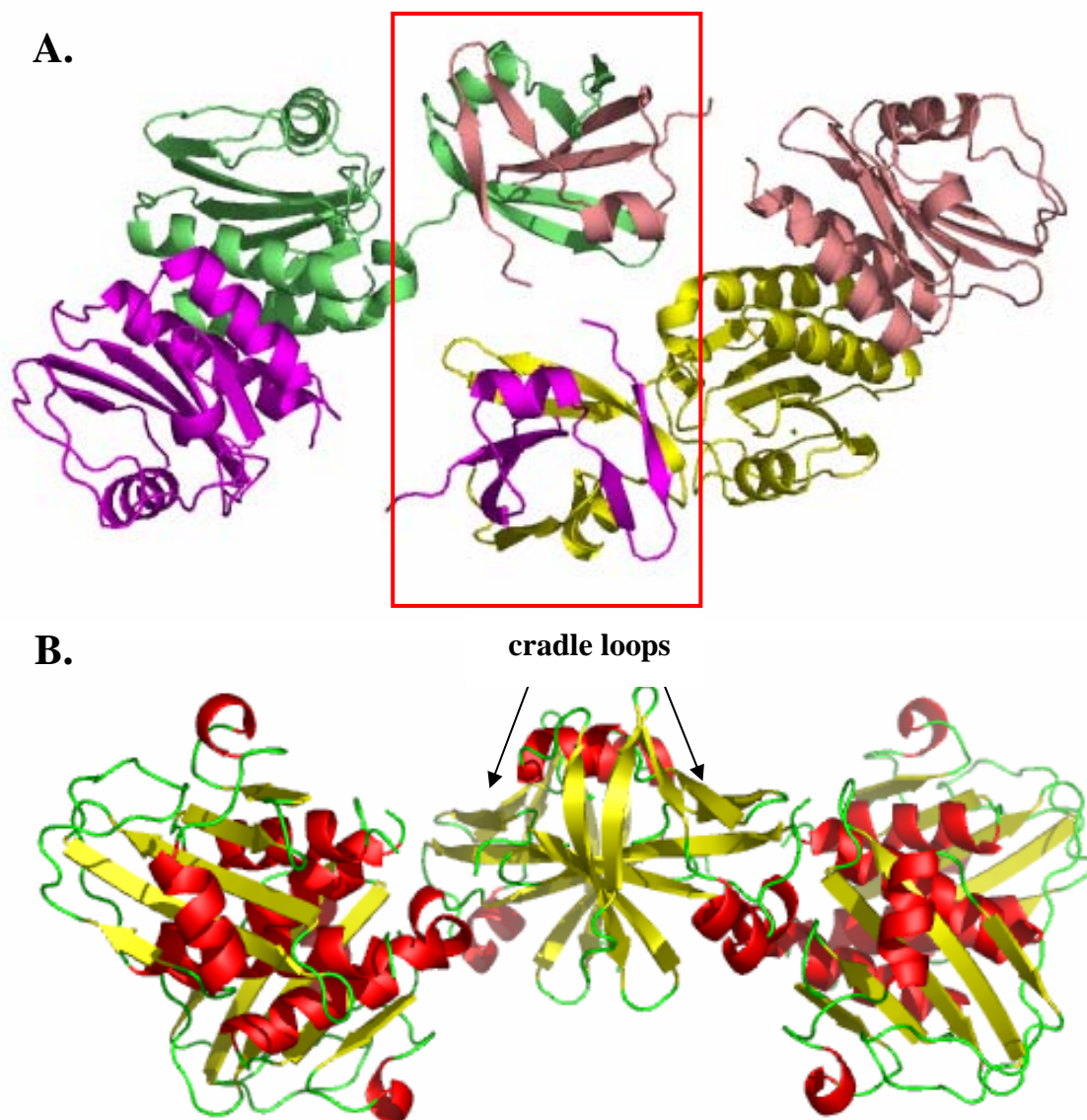
Figure 4.23(A.) Structure of the SpoVT tetramer in top view; each chain is colored differently; N-terminal domains that form the AbrB-N like cradle-loop barrels are highlighted by a red box. **(B.)** SpoVT tetramer structure in side view; black arrows indicate the position of the so called cradle-loops which are tilted by approximately 45° in respect to each other.

Partial proteolysis of the SpoVT protein was performed in the presence of possible GAF binding ligands using 0.26µg of trypsin. Samples were incubated at 25 °C in 0.12 ml of buffer containing 20 mM TRIS/HCl, pH 8.0, 1 mM $MgCl_2$, 1mM dithiothreitol, and 0.5 mM EDTA. At various time points, 10 µl samples were removed, denatured, and subjected to SDS-PAGE (Fig. 4.24). Binding was tested with cGMP, GTP, cAMP, ATP, Leu, Ile and Val, at concentration of 10mM. The same

experiments were repeated using chymotrypsin (0.52μg) and proteinase K (0.4μg) for proteolysis. A different pattern as well as a change in overall stability in the presence of protease was observed only in experiments where GTP was added to the protein solution would indicate that GTP is the ligand for GAF domain of SpoVT (Fig. 4.24A). Control with BSA had same patterns whether the GTP was present or not in mixture (data not shown).
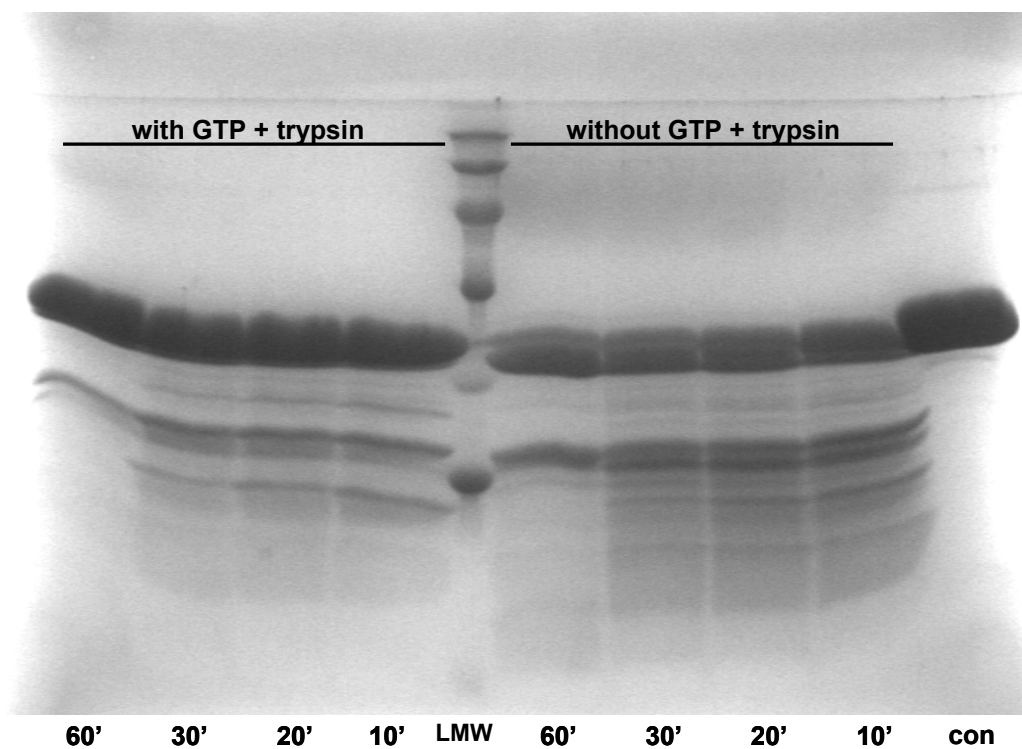
Figure 4.24 Partial proteolysis of the SpoVT protein with trypsin. On the bar above lines is indicated in which samples 5 mM GTP was added; time points of aliquots are indicated under the figure. The lane on the far right side represents protein control (con) without added trypsin and the lane in the middle is the LMW marker.

4.1.11 Bioinformatic analysis of Mj0056

A PSI-BLAST search using Mj0056 (gi 2495770) as the query sequence was performed and resulted in two distinct sets of sequences. Both sets of sequences showed similarity with one canonical half of the VAT-Nn barrel at their C-termini, while the N-terminal part that was alignable with any sequence. The first set of proteins was of similar size as Mj0056 (120-150 residues, blue box in Fig. 4.25), while the second group of protein sequences had relatively longer sequences (200-250 residues, red box in Fig. 4.25).

```
                                      repeat
                          |---------------------------------------------|
                            β1          Ψ        β2   α1              β3
                            sssssss            ssss hhhhh          ssssss
VAT-Nn Ta        mesnnGIILRVAEANSTDPG-MSRVRLDESSRRLLDAE-IGDVVEIEKVr
                 KTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV            90
                            sssss              ssss  hhhhhh          ssssss
Af2106          [190] EGAIVIPKRTHYPAE-ILEVISPVKLRDKLGLK-DGDFVEVEVIl
        233
Ta1064          [178] PCFVIMPERTVYTD--VIEIISDKYLREEINLH-DGDRVSVEVYt
        220
Ph0660          [170] EGAIVIPSRTVHPPK-IAEVIAPVNLRKTLNLQ-DGDKVRIKVL      212

Mj0056          [ 90] DGAIVVPKKTYHSSE-IIEIIAPMKLREQFNLK-DGDVIKILIKgdkde 136
Mt1500          [ 75] EGAILFPLKTHHRQG-CLEFVAPVNLRKTLKLR-DGDTVSLDIDtseiqe 122
```

Figure 4.25 Alignment of VAT-Nn with Mj0056 and related sequences. The two sets of sequences are labeled with different colored boxes. Residues forming the hydrophobic core of the barrel are colored blue; GD box and positive charges are colored red.

An analysis with "The SEED" using different sequences from these groups as starting points revealed, that the majority of genes for Mj0056-like proteins are connected with the gene for 3,4-dihydroxy-2-butanone-4-phosphate synthase, an enzyme of the riboflavin biosynthesis pathway (Fig 4.26). Therefore, Mj0056-like proteins are annotated as transcriptional regulators of the riboflavin synthesis pathway in most sequenced genomes. Conservation of the gene loci and gene orientation as well as the short distance between the two genes (10-30 bp) are in favor of such an annotation, but no experimental data are available to support this hypothesis.
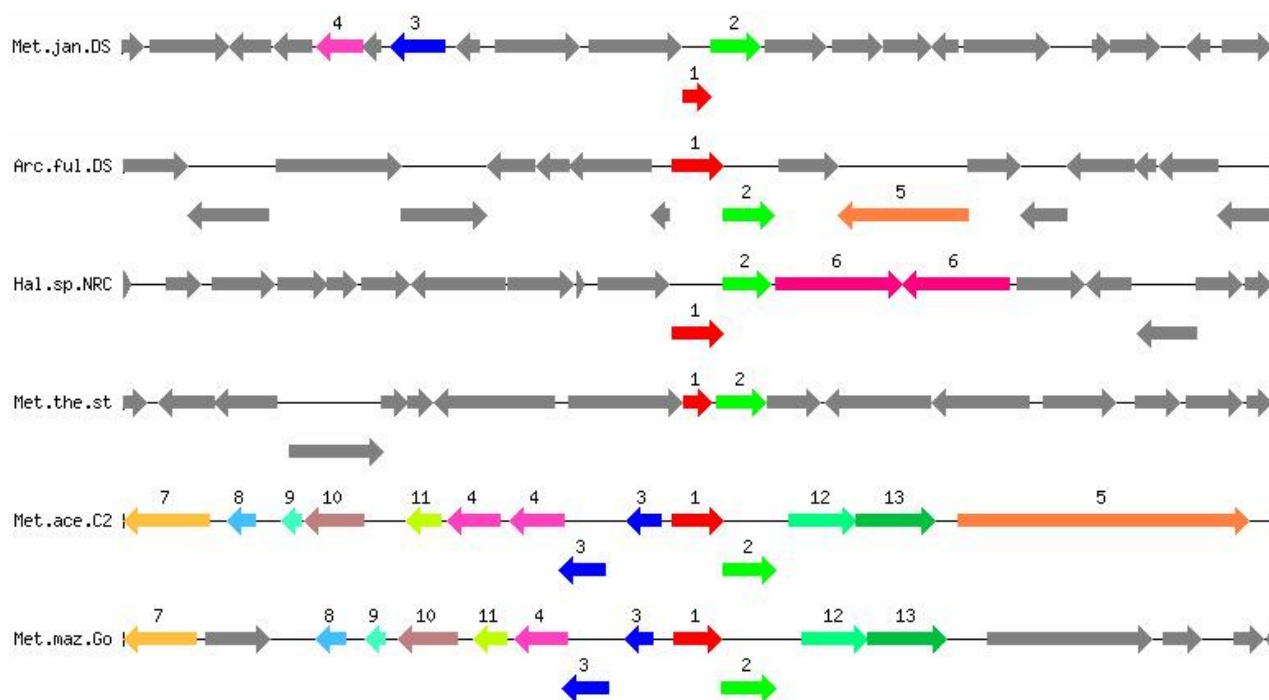
Figure 4.18 Analysis of the Mj0056 gene locus using "The SEED"; Mj0056-like genes are labeled with red color and number 1; the gene labeled by green color and number 2 encodes the 3, 4-dihydroxy-2-butanone 4-phosphate synthase, an enzyme from the riboflavin biosynthesis pathway. Genomes are (from top to bottom): *Methanocaldococcus jannaschii*, *Archaeoglobus fulgidus*, *Halobacterium sp.*, *Methanobacterium thermoautotropicum*, *Methanosarcina mazei* and *Methanosarcina acetivorans*.

4.1.8 NMR structure of Mj0056

Mj0056 was amplified from genomic DNA of *Methanocaldococcus jannaschii* by polymerase chain reaction (PCR) and cloned into the pET30b vector (Novagen). For expression in *E. coli C41*, cells were grown in LB medium at 37°C, induced at OD_{600}~0.6 with 1 mM IPTG and harvested after 4h. The protein was purified using a combination of MonoQ HR 5/5 (GE Healthcare) anion exchange and SP-Sepharose Fast Flow (GE Healthcare) cation exchange columns. The final step of the purification included buffer exchange by gel filtration on a Superdex G-75 column (GE Healthcare). The protein was concentrated to 10mg/ml and prepared for

crystallization in 20 mM MOPS, pH 7.25, 120 mM NaCl. Crystallization was done by the hanging drop method using a 1:1 volume ratio of protein to precipitant solution with Hampton research plates and crystal screen solutions. Crystals of similar shape were noticed in two drops, both containing 2.0 M $(NH_4)_2SO_4$ as precipitant (Fig. 4.26).
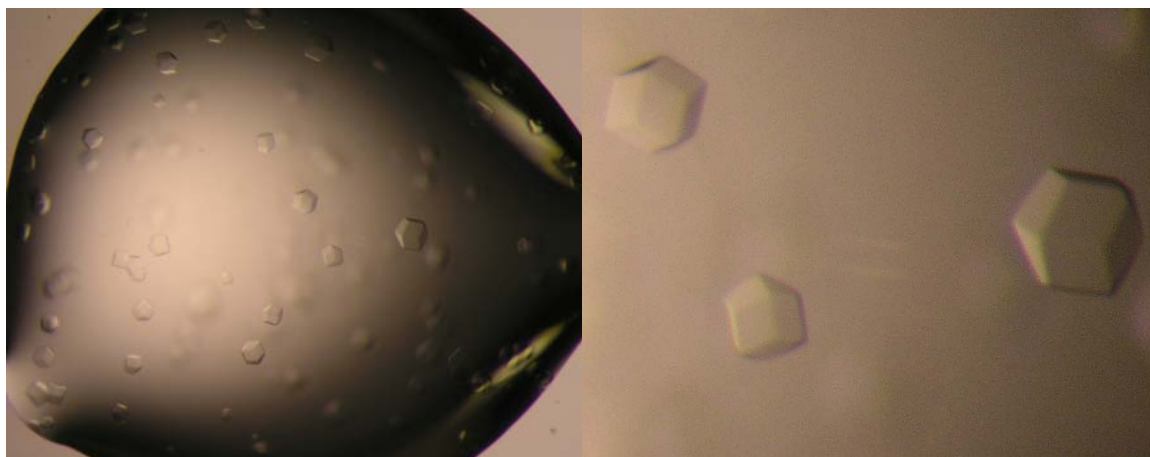
Figure 4.26 Crystals of the Mj0056 protein (c = 10 mg/ml, 0.1 M Na-acetate, pH 4.6, precipitant 2.0 M $(NH_4)_2SO_4$)

Crystals of the Mj0056 were diffracting to a resolution of 6 Å and Se-Met derivates improved the resolution to 3.3 Å, but even this resolution was too low for structure calculation, since no good model could be built for molecular replacement. The protein structure was solved using NMR spectroscopy. Uniformly ^{15}N- or ^{13}C-labeled Mj0056 protein was produced by growing bacteria in M9 minimal medium with ^{15}NH$_4$Cl (0.7 g/l) and ^{13}C$_6$-glucose (2 g/l) as sole nitrogen or carbon sources. The protein was purified by the same method described above, with the exception that the final buffer was Na-phosphate, pH 7.4, 150 mM NaCl. The structure of Mj0056 (Fig. 4.27) was solved by Murray Coles and Vincent Truffault, but was not yet refined and submitted to the RSCB Protein Data Bank (http://www.rcsb.org/pdb). The NMR structure of the Mj0056 protein will be used like a model for molecular replacement and calculation of the crystal structure. In such way we will check for possible presence of the ligand that could not be seen in the NMR structure.
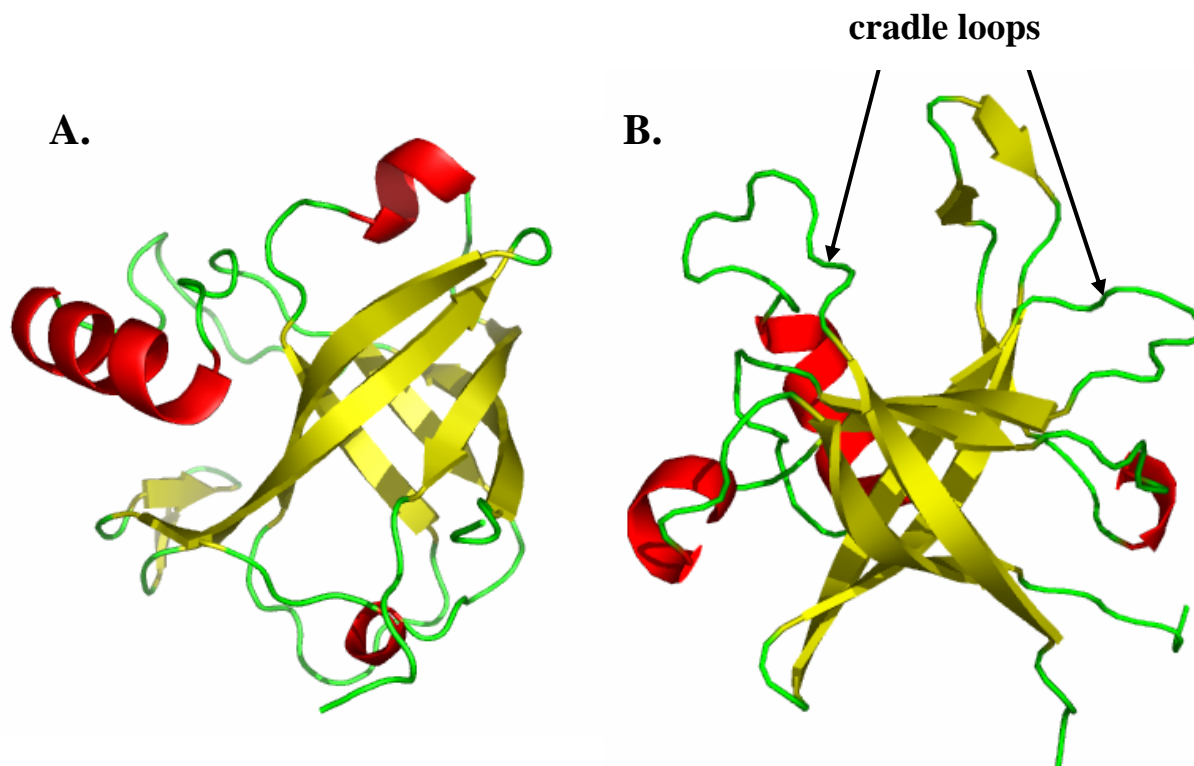
Figure 4.27(A.) NMR structure of the Mj0056 protein (side view); **(B.)** Top view on structure of the Mj0056 protein; black arrows indicate position of the loops that are analogous to the psi-loops of VAT-Nn or the cradle-loops of AbrB-N

The structure of Mj0056 resembles a 6-stranded beta barrel fold capped from both sides with small helices. The overall structure of the barrel is very similar to the structure of the Phs018 protein (Coles et al., in preparation) with insertions of additional secondary structure elements in several loops. The topology and connection of the Mj0056 structure with structures of other members of double-psi, swapped-hairpin and RIFT barrels is discussed later.

4.2 β-clam domains

Phylogenetic analysis of the AAA family of proteins revealed a new subgroup of AAA proteins which are in most of databases designated to be FtsH or cell division control AAA proteins. We have named this subgroup of AAA proteins AMA, since they can be found only in *Archeglobus f.* and the methanogenic group of archaeabacteria. Bioinformatic analysis of the N-terminal domain of AMA proteins showed that they could assume a CDC48 domain 2-like fold, a six stranded clam-like structure. This fold is found in the N-terminal part of CDC48-like proteins and also in UFD1 proteins. In those two groups of proteins this domain is following another domain (double psi-barrel fold) forming a kidney-shaped structure. It has been shown that the combination of the two named domains, like in VAT (Cdc48/p97 homolog from *Thermoplasma acidophilum*) and yeast UFD1 play a role in binding protein substrates (Golbik et al., 1999; Park et al., 2005). Characteristic of AMA proteins is that a clam-like domain is the sole N-terminal domain. This prompted us to check whether the N-terminal domain of AMA proteins can still bind protein substrates and prevent their aggregation, since it is missing a double psi-barrel domain.

4.2.2 Expression and purification of AMA constructs

DNA sequences encoding AfAMA (Af1285) and its N-terminal and ATPase domains, AfAMA-N (residues 1-101) and AfAMA-ΔN (residues 102-352), respectively, were amplified by PCR from genomic DNA of *A. fulgidus*. The Mj-AMA (Mj1494) and MjAMA-N (residues 1-120) sequence were amplified by PCR from genomic DNA of *M. jannaschii*. Constructs were cloned into the pET-30b expression vector (Novagen) via the *Nde*I and *Hind*III restriction sites. All constructs were generated as N-terminal 6xHis-tag fusions to facilitate protein purification, but if required were recloned without tag (AfAMA-N, MjAMA-N). AfAMA-N was also cloned into the pThiohis(b) vector (Invitrogen) as a C-terminal fusion to a modified thioredoxin (Thio-AfAMA-N), using *Nco*I and *Xba*I restriction sites. The identity of all constructs

was confirmed by DNA sequencing. All proteins were expressed in *E. coli* C41 (DE3) at 37°C after induction with 1mM IPTG at OD_{600}~0.6. With the exception of AfAMA-N, all His6-tagged proteins remained in the soluble fraction of cellular extracts and were purified by a combination of immobilized metal affinity and gel-size exclusion chromatography. The majority of AfAMA-N was found in inclusion bodies. This fraction was solubilized in low-salt buffer (50mM Tris/Hcl, pH 7.3, 50mM NaCl) containing 8M urea, and purified by ion-exchange chromatography under denaturing conditions with 8M urea. 95% of the protein was refolded by overnight dialysis against PBS, 1mM DTT and 5% glycerol. AfAMA-N from the soluble fraction of the lysate was purified first by ion-exchange chromatography. Samples from both purification procedures were finally purified by gel-size exclusion chromatography. Protein purity was >95% as judged by SDS-PAGE. Both AfAMA-N protein preparations, soluble and refolded, behaved identically and had a very low solubility (0,4 mg/ml) in the absence of 5% glycerol. To avoid precipitation, we expressed AfAMA-N fused to the C-terminus of thioredoxin (Thio-AfAMA-N) (Hammarstrom et al., 2002). Indeed, solubility of the fusion protein was 40-50 times increased, but removal of thioredoxin by cleavage with Factor Xa (Novagen) led again to precipitation. To obtain information about the molecular mass and oligomeric state of all proteins, samples were analyzed on analytical Superose 6 and 12 columns (GE Healthcare), calibrated with gel filtration standards (Sigma) and additional standard proteins. Since oligomerization of some AAA proteins is ATP dependent, samples were also analyzed in the presence of 5 mM ATP (Fig. 4.28 and 4.29).

The retention time of AfAMA was between the two size markers thyroglobulin (669 kD) and ferritin (474 kD), and based on a monomeric MW of 40 kD for AfAMA, this suggests a dodecameric complex of approx. 480 kD., AfAMA-ΔN (29 kD), which comprises of the AMA ATPase domain, did not form higher oligomers but was eluted as a dimer on gel-size exclusion columns. The presence of ATP did not change the oligomerization state.
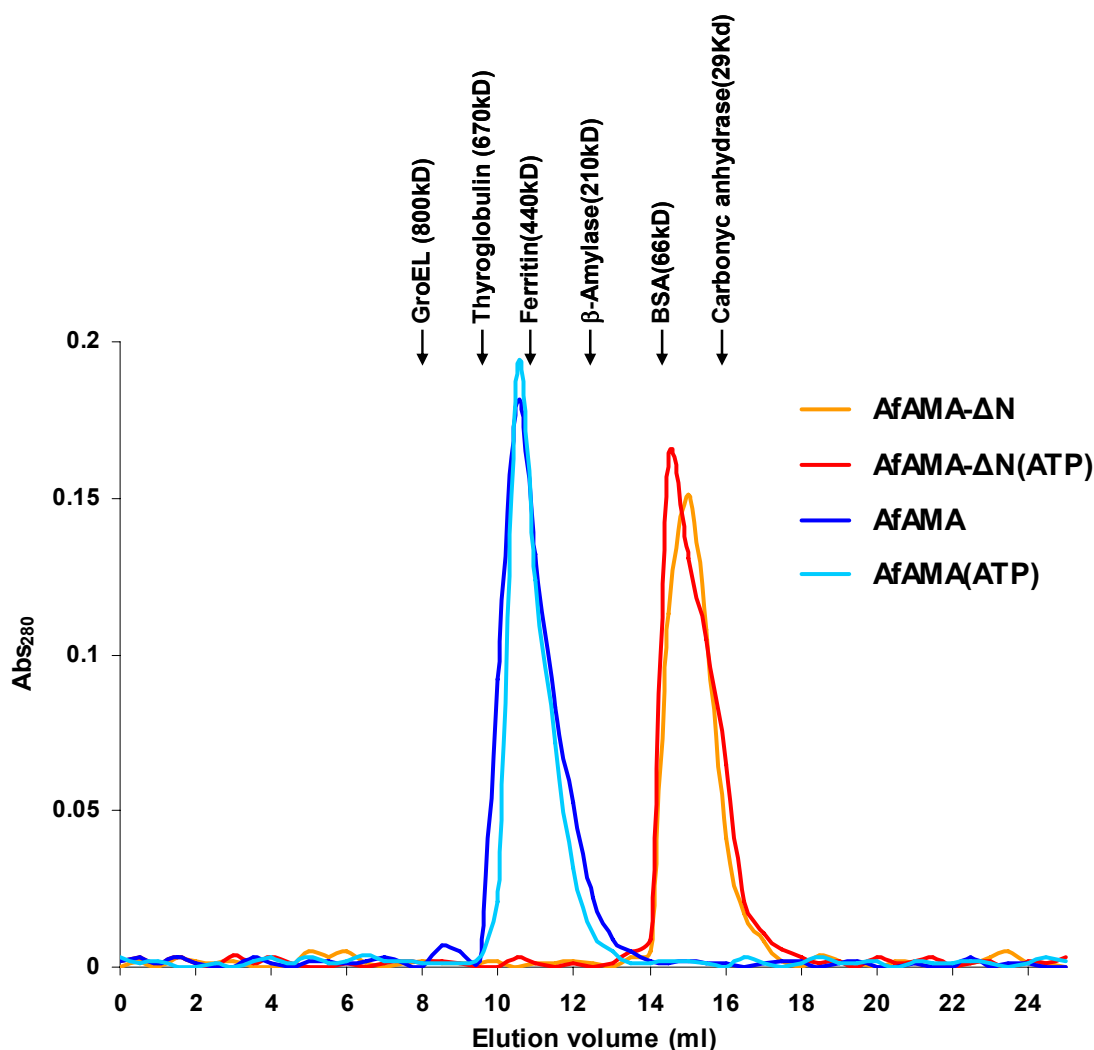
Figure 4.28 Elution profiles of AfAMA and AfAMA-ΔN in the absence or presence of nucleotide (ATP) on a Superose 6 gel-size exclusion column. Arrows above the elution profiles indicate the elution volumes of proteins used for calibration of the column.

As the oligomerization of AfAMA was obviously a consequence of its N-terminal domain AfAMA-N, we determined also the native molecular weight for this domain (monomeric size 12 kD). Based on elution on the Superose 12 gel-size exclusion column the size is approx. 70 kD, corresponding to a hexameric complex (Fig 4.29). Likewise, Thio-AfAMA-N (monomeric size 26 kD) and MjAMA-N, the homologous N-domain from *Methanocaldococcus jannaschii*, migrated as hexamers on the gel filtration column (data not shown).
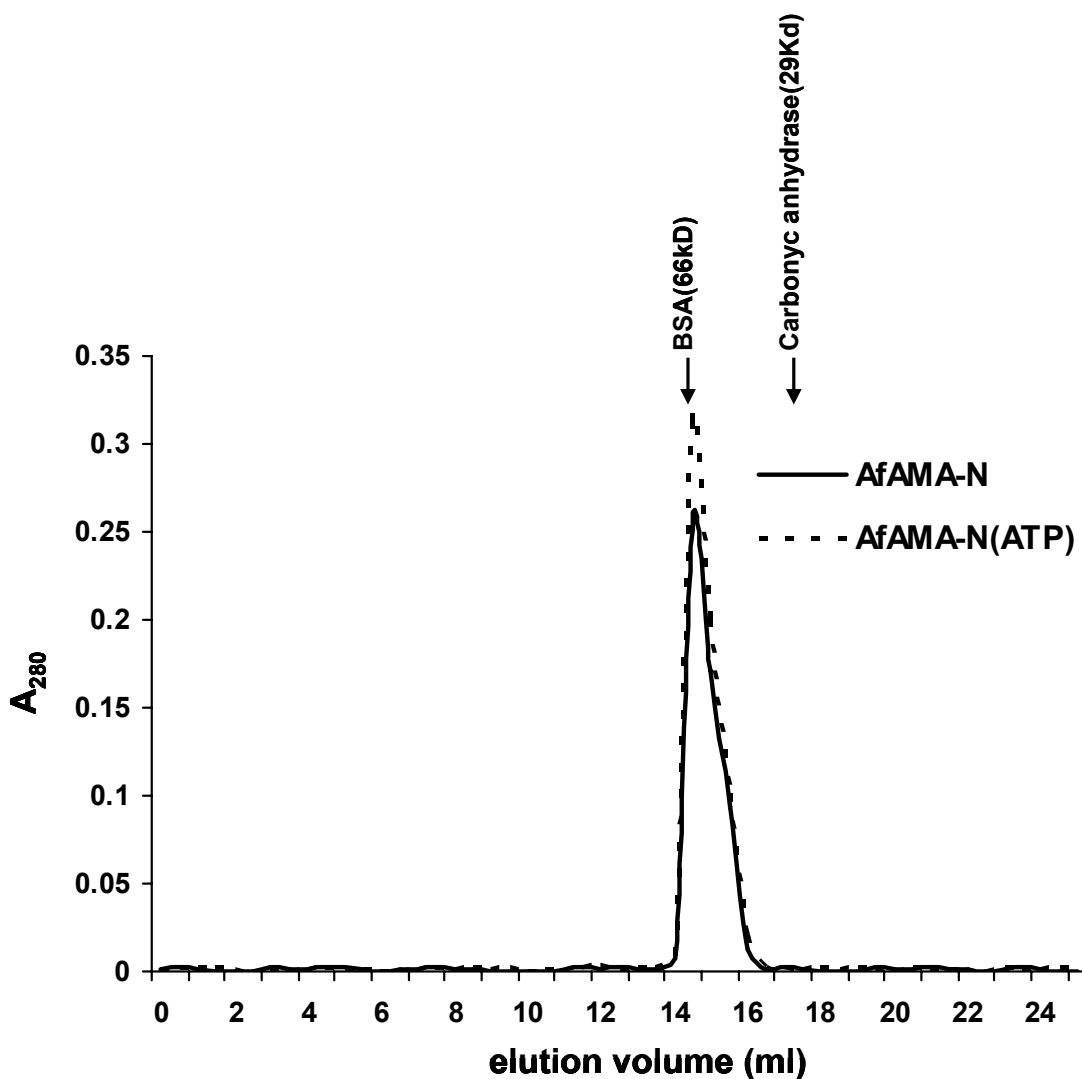
Figure 4.29 Elution profiles of AfAMA-N in the absence and presence of nucleotide (ATP) on a Superose 6 gel-size exclusion column. Arrows above the elution profiles indicate the elution volumes of proteins used for calibration of the column. To put emphasis on molecular weight only two molecular weight markers are shown (BSA and carbonic anhydrase)

CD spectra of all constructs indicated folded proteins with a mixture of α-helix and β-strand content for AfAMA and AfAMA-ΔN (minima at 222 and 209 nm), and mainly β-sheet content for AfAMA-N (minimum at 216 nm) (Fig. 4.30).
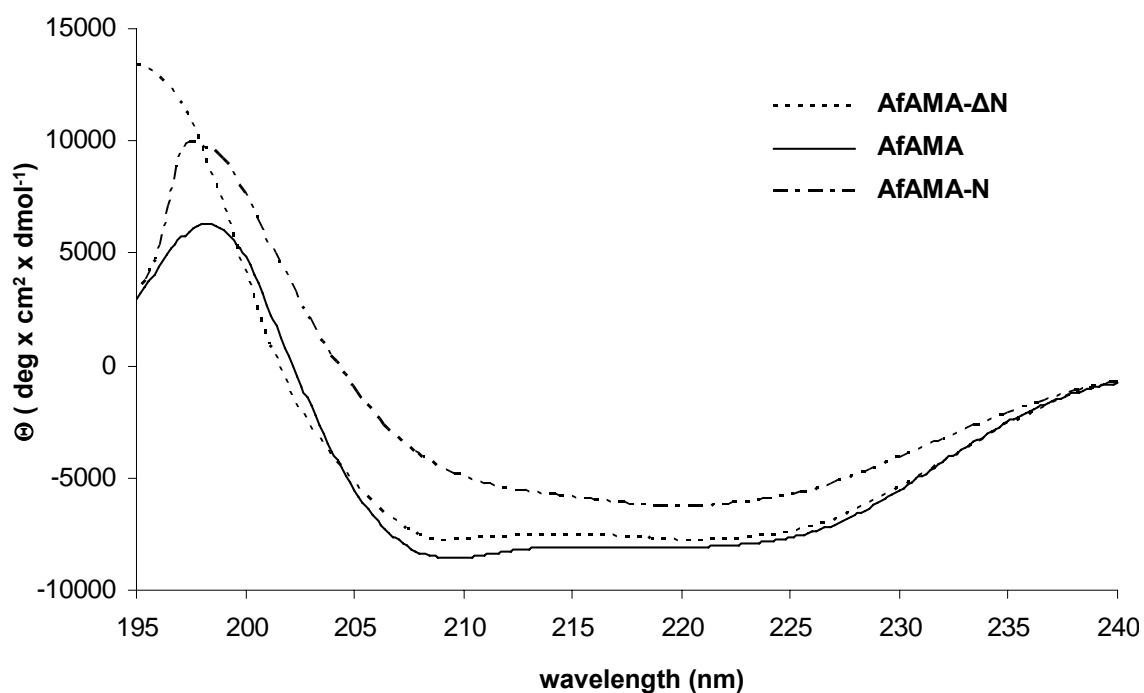
Figure 4.30 CD spectra of recombinant AfAMA, AfAMA-ΔN and AfAMA-N
proteins

Thermal denaturation curves revealed melting temperatures (Tm) for AfAMA-ΔN and AfAMA of 84±1°C and 88±1°C, respectively (Fig 4.31). The sigmoidal shape of the melting curves indicates that the unfolding process in both proteins is cooperative, excluding a folding defect as reason for the behavior of AfAMA-ΔN. The lower Tm for AfAMA-ΔN is probably due to the missing N-terminal domain.

Tm for AfAMA-N could not be determined by this method, as the CD spectrum did not show any obvious changes up to 98°C (data not shown). This indicates a very high stability of the isolated N-domain complex. CD spectra recorded after boiling the protein 5 minutes at 100°C did not show any changes to the one shown in Figure 4.31. This stability is further substantiated by the finding that AfAMA-N is stable in SDS-sample buffer, even after 5 min of boiling at 95°C (Fig. 4.32).
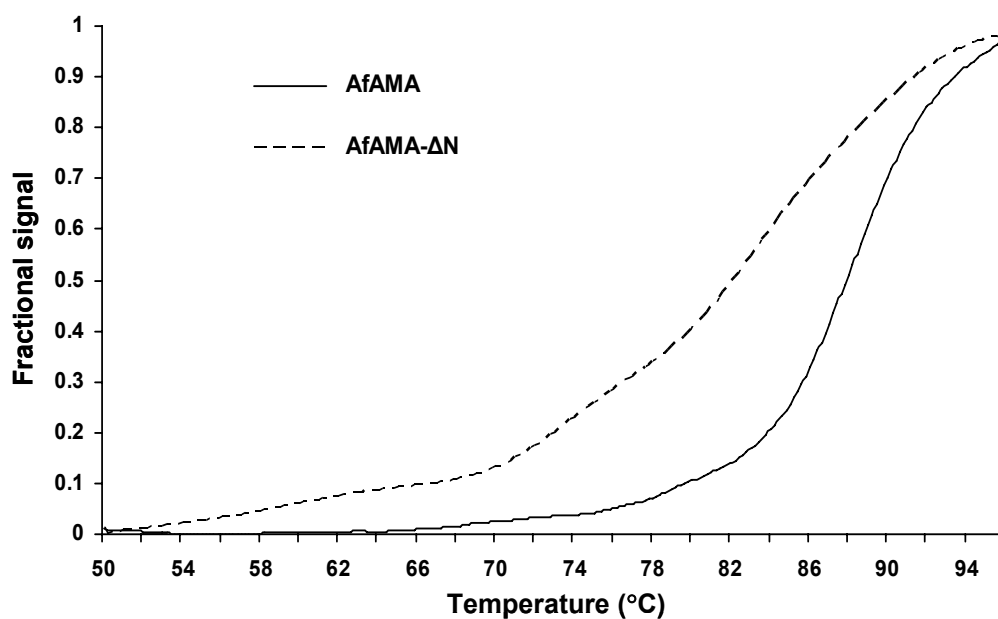
Figure 4.31 CD heat-denaturation curves of AfAMA and AfAMA-ΔN. The fractional signal represents an arbitrary scale based on difference of signals of folded and unfolded state of protein (or signal at room temperature and 98°C)
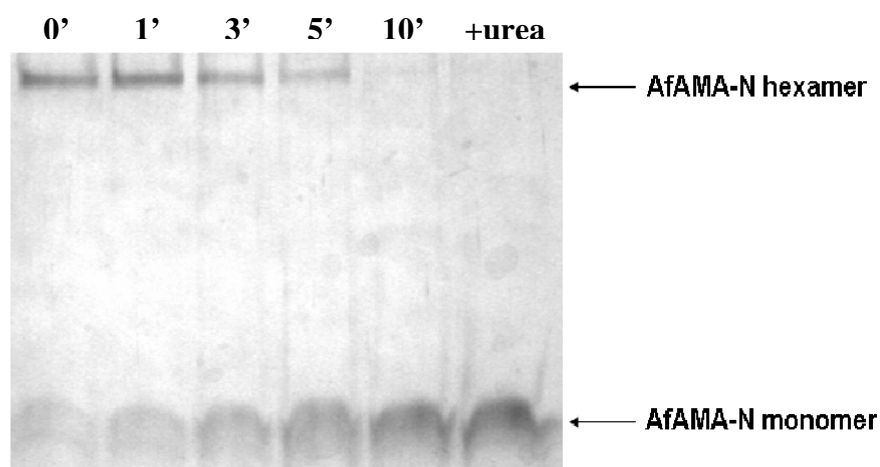
Figure 4.32 SDS-PAGE gel of hexameric and monomeric AfAMA-N after incubation in SDS-containing buffer at 95°C. When indicated, 6 M urea was added to the sample

4.2.2 Electron microscopy of the AMA constructs

Images of AfAMA-N complexes were recorded using electron microscopy (EM) at the in-house EM service facility with the help of Dr. Heinz Schwarz. When adsorbed to carbon-coated grids, negatively stained AfAMA-N particles appeared as ring structures with a stain-filled center (Figure 4.33). Interestingly, the addition of a 6xHis tag to AfAMA-N resulted in a totally different appearance of the sample in EM. The formation of elongated, chain-like structures was observed (data not shown).
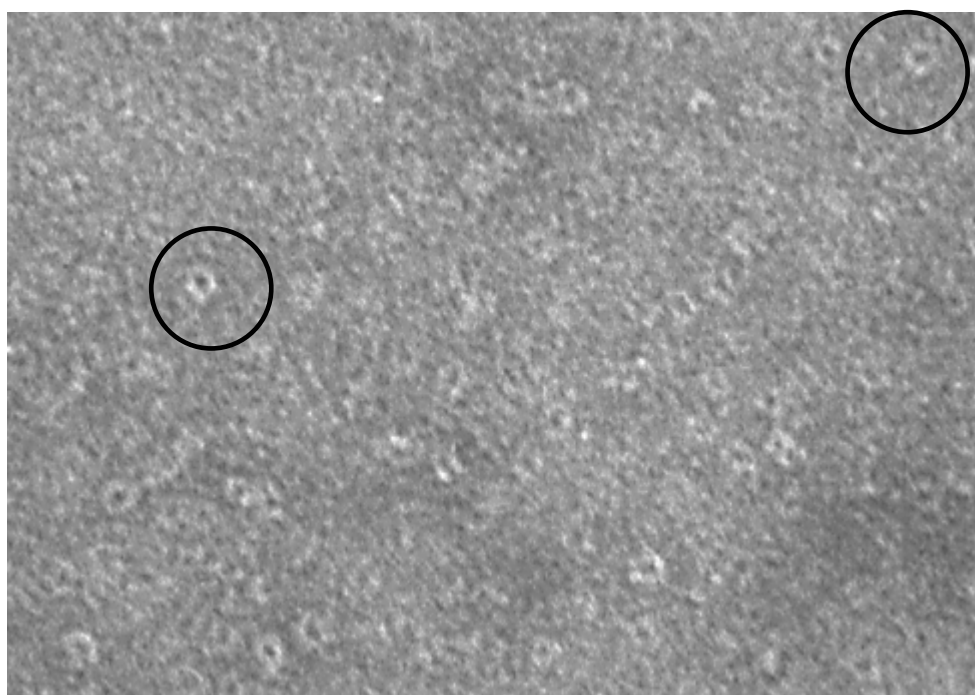
Figure 4.33 AfAMA-N particles negatively stained in uranyl acetate. Some of the ring structures are encircled.

Averaging of negatively stained AfAMA-N particles was done in collaboration with Dr. Beate Rockel from MPI of Biochemistry in Martinsried. Averages of the AfAMA-N particles display hexameric symmetry, which is obvious from the eigenvectors obtained from the analysis of a particle stack. Those particle images were centered and afterwards corrected for their rotation (Fig. 4.34A, B). The diameter of the hexameric class averages is approx. 9 nm with a stain-filled centre of approx. 1.5 nm in diameter.
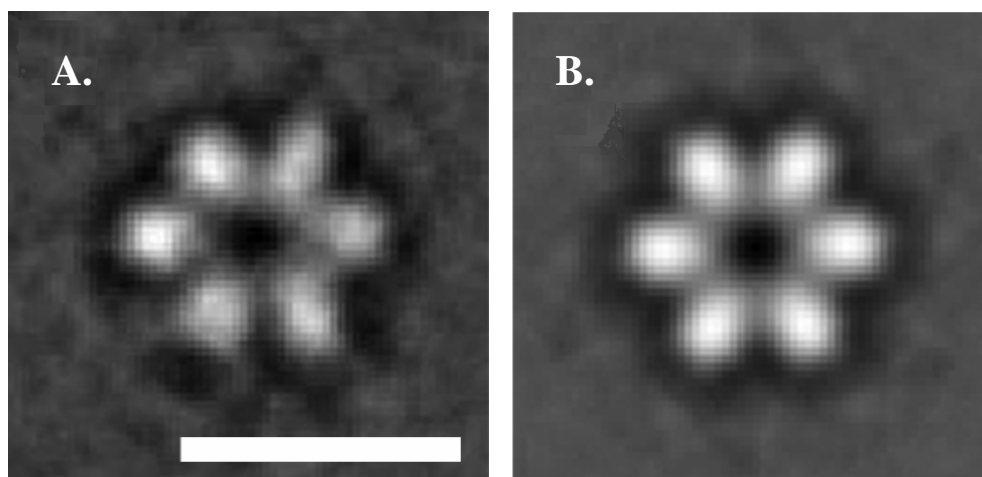
Figure 4.34(A.) Class average of AfAMA-N representing hexameric top views. The scale bar represents 10 nm. **(B.)** Sixfold symmetrized class average.

Images of the AfAMA complex could not be obtained at good resolution. Ring structures were visible but the overall shape of the particles appeared to be non-uniform, which made averaging of the images impossible. Addition of glycerol (5-10 % v/v) to the sample buffer improved the behavior of the complex (data not shown).

4.2.3 Chaperone activity of AMA constructs

The physiological role of AMA proteins is unknown. Yet, it is reasonable to assume that AMA proteins, like other AAA ATPases, are involved in protein unfolding and the dissociation of complexes, possibly in cooperation with other chaperones or proteases. Therefore AfAMA was tested for its ability to interact with non-native proteins, as they may also be presented during unfolding of physiological substrate proteins. These artificial substrates were generated by heat-induced denaturation of two model proteins, firefly luciferase (Fig. 4.35) and porcine citrate synthase (Fig. 4.36). The resulting aggregation of these proteins was measured as increase in attenuance caused by light scattering.
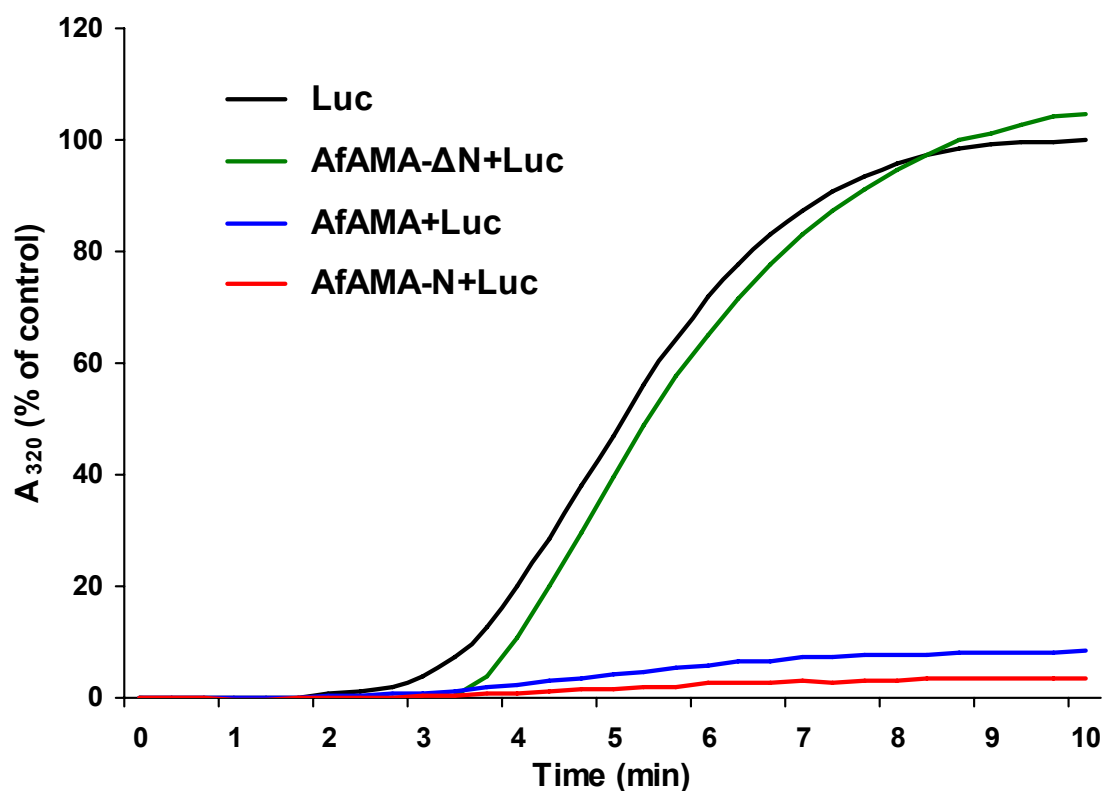
Figure 4.35 Chaperone activity of AMA proteins. Heat-inducible aggregation of luciferase at 43 °C is measured in the presence of equimolar concentrations of AfAMA, AfAMA-ΔN, or AfAMA-N.

When AfAMA or AfAMA-N were present in stoichiometric concentrations during the heat-incubation period, aggregation of both citrate synthase and luciferase was suppressed, demonstrating an intrinsic chaperone activity of AfAMA, which resides in the N-terminal domain. AfAMA-ΔN had no effect in these assays, indicating that either the ATPase domain of AMA is not involved in substrate binding, at least with the proteins we tested, or that it requires the hexameric ring structure to make stable contacts with non-native proteins.
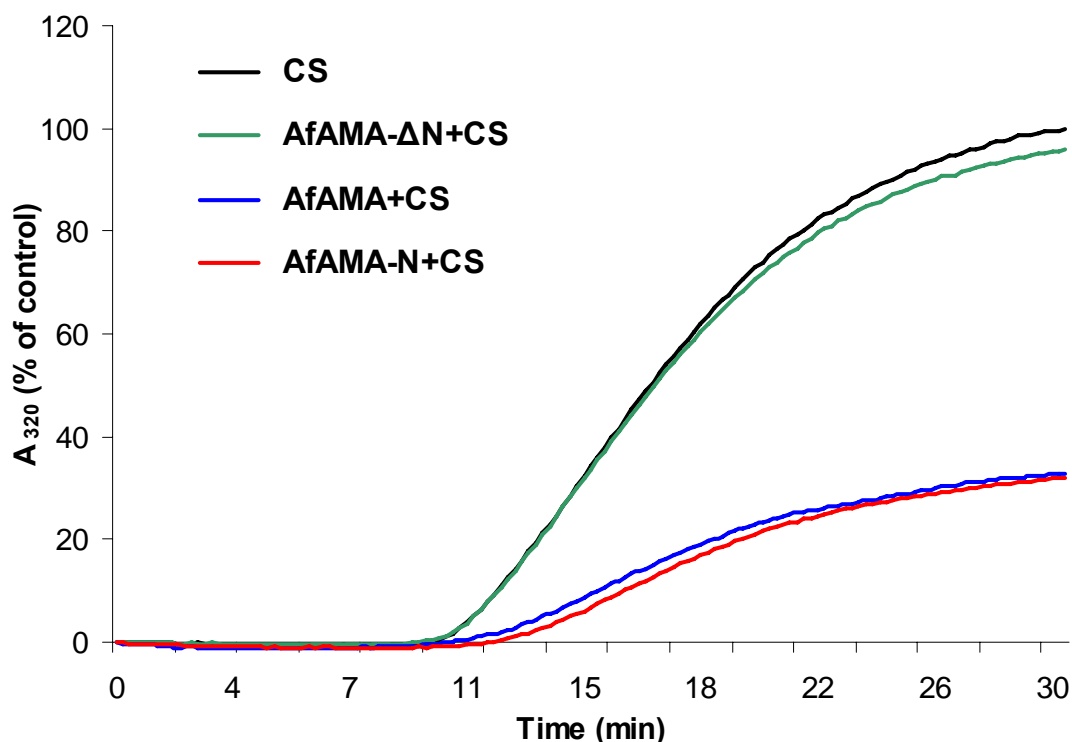
Figure 4.36 Chaperone activity of AMA proteins. Heat-inducible aggregation of citrate synthase at 50 °C in the absence or presence of a threefold molar excess of AfAMA, AfAMA-ΔN or AfAMA-N.

According to the generic reaction mechanism of AAA ATPases, the interaction of AfAMA with non-native proteins should be affected by available nucleotides for hydrolysis. Indeed, when ATP was present in the reaction mixture from the beginning, AfAMA did not suppress aggregation anymore, suggesting that substrate protein was released immediately after binding or was not bound in the first place (Fig. 4.37). Similarly, release of substrate protein could be induced by addition of ATP at later time points (Fig. 4.37). In the latter case, the concerted release of proteins by addition of ATP is expected to generate at once high concentrations of unfolded polypeptides in solution, resulting in very rapid aggregation. Non-hydrolyzable analogs ATPγS and AMP-PNP were not effective in dissociating substrate protein (Fig. 4.37), indicating that hydrolysis rather than mere binding of ATP is required to elicit this step. Nucleotides had no effect on the interaction of proteins with AfAMA-N (not shown).
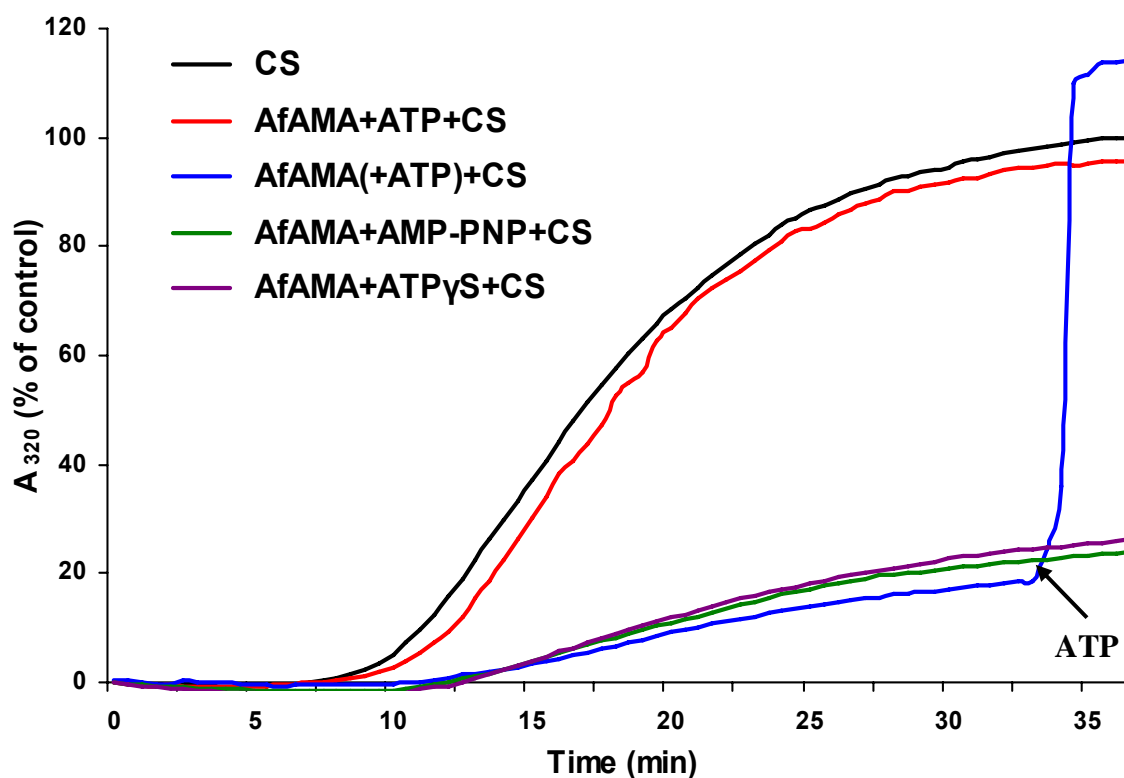
Figure 4.37 Citrate synthase aggregation in the absence or presence of a fivefold molar excess of AfAMA. To test for effects of nucleotides, ATP, ATPγS or AMP-PNP were either present from the beginning or, in case of ATP, added at a later time-point (indicated by arrow).

4.2.4 Temperature dependant ATPase activity of AMA constructs

A defining feature of the AAA protein family is their ATPase activity, which serves to dissociate bound substrates and to switch the complex between high- and low-affinity states. Given the thermostable nature of AfAMA, I assayed all constructs for ATPase activity at high temperatures (Fig. 4.38). AfAMA showed maximal ATPase activity close to the optimal growth temperature of *A. fulgidus* (80°C), but had no activity at temperatures below 45°C. Interestingly, AfAMA-ΔN had an even more pronounced ATPase activity, with a maximum at 75°C (Fig. 4.38).
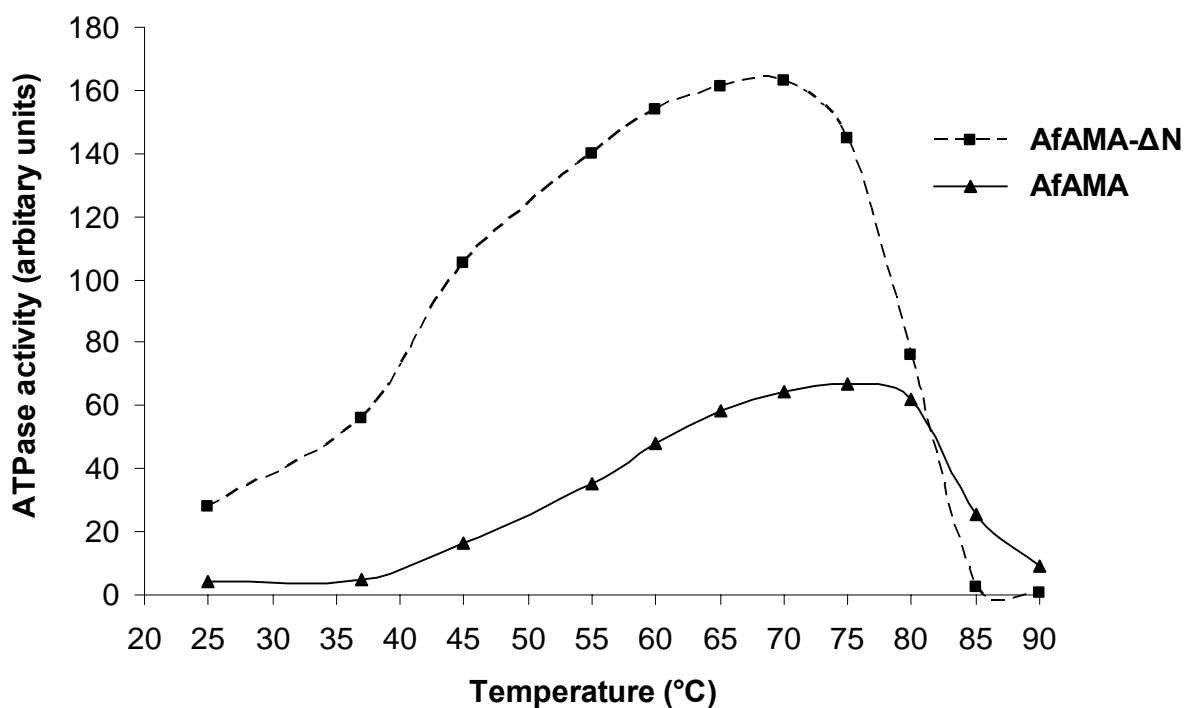
Figure 4.38 ATPase activity of AfAMA (▲) and AfAMA-ΔN (■) measured at different temperatures.

In fact, the specific activity of AfAMA-ΔN was approximately three-fold higher than that of full-length AfAMA (~990 nmol mg^{-1} min^{-1} vs. ~380 nmol mg^{-1} min^{-1} at 70°C), the latter one being in the same range as the ATPase activity of FtsH at 37°C (~450 nmol mg^{-1} min^{-1}) (Karata et al., 1999). As expected, AfAMA-N did not have any ATPase activity (data not shown).

4.2.5 GYPL and deletion mutants of AMA

Sequence similarity searches using BLAST or PSI-BLAST with the AMA-N domain like a query did not yield matches outside the AMA group. Nevertheless, the similarity of this domain to the β-clam part of the N-domain found in proteins of the Cdc48/p97 clade (Appendix Fig. 4) is readily apparent with more advanced methods, such as HHsenser, which uses Hidden Markov Model comparisons (Söding, 2004). The alignment of these domains to their homolog from *T. acidophilum* (VAT-Nc) is shown in Figure 4.39. Besides the relatively longer loop region between the first β-strand and α-helix, a characteristic of the AMA-N sequences is the presence of a conserved GYPL motif in the same loop. To elucidate the importance of this motif using standard site-directed mutagenesis techniques, I constructed the AfAMA mutants Y19A, L21A, double mutant Y19A; L21A as well as a deletion mutant AMA-N-ΔGYPL. Constructs were cloned into the pET30b expression vector (Novagen) with *Nde* I and *Hind*III restriction sites. All constructs were generated as N-terminal 6xHis-tag fusions to facilitate protein purification. Proteins were expressed in *E. coli* C41 (DE3) at 37°C after induction with 1mM IPTG at OD_{600}~0.6. With the exception of AMA-N-ΔGYPL all His6-tagged proteins remained in the soluble fraction of cellular extracts and were purified by a combination of immobilized metal affinity and gel-size exclusion chromatography. Refolding of the AMA-N-ΔGYPL was not possible under any of the tested conditions. Mutant proteins were analyzed on analytical Superose 6 and 12 columns (GE Healthcare), calibrated with gel filtration standards (Sigma) and additional standard proteins to obtain information about their oligomerization. It was found that AfAMA requires an intact GYPL motif for hexamerization, as single or double mutations in the conserved GYPL motif resulted in the loss of the oligomeric ring structure (Fig. 4.40, mutant data are shown only for AfAMA L21A).

A.

B.

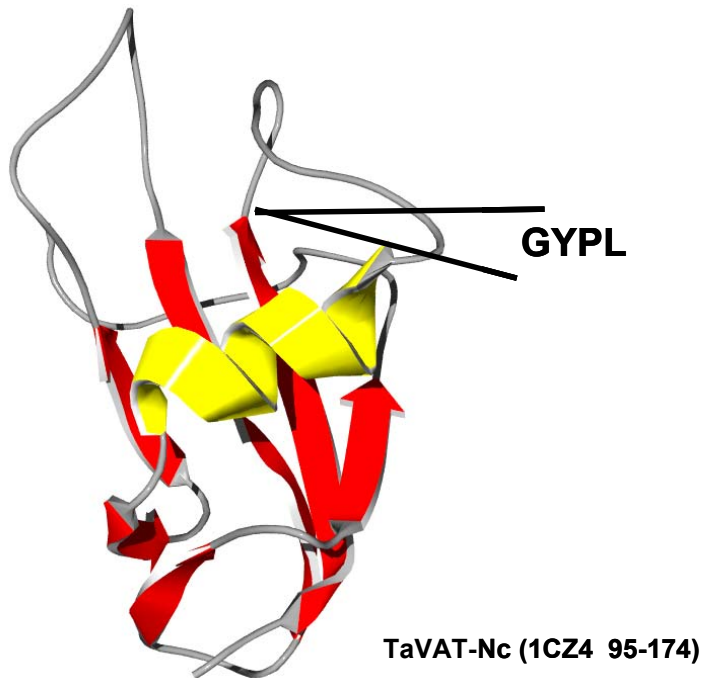**GYPL**

TaVAT-Nc (1CZ4 95-174)

Figure 4.39 (A.) Alignment of AMA-N sequences with the homologous VAT-Nc domain. Red and yellow boxed residues are forming β-strands and α-helix, respectively. The blue box highlights the position of the GYPL motif in the AMA-N domains, which is absent in VAT-Nc.

(B.) Structure of the VAT-Nc domain (1CZ4, residues 95-174). The arrow indicates the position of the GYPL motif in homologous AMA-N domains
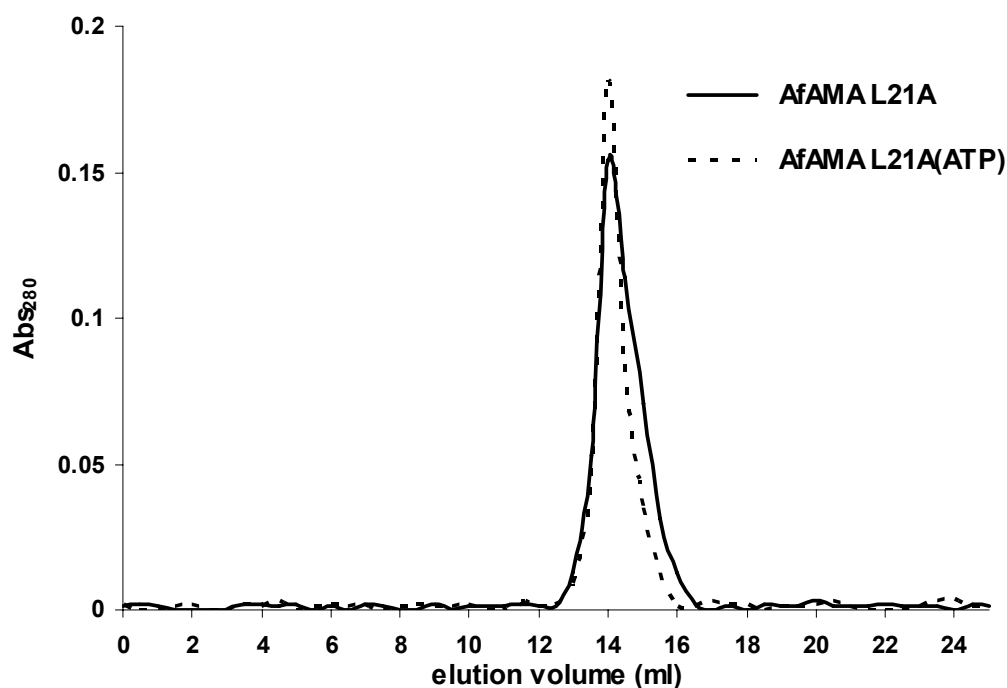
Figure 4.40 Elution profiles of AfAMA L21A mutant in the absence or presence of nucleotide (ATP) on a Superose 6 gel-size exclusion column.

To check whether the presence of nucleotides can favors oligomerization, column runs for mutant proteins were repeated in the presence of ATP. However, there was no change in any of the chromatograms (shown exemplary for the L21A mutant in figure 4.40), showing that AfAMA mutants are dimers also under these conditions, and lost their ability to assemble into hexameric rings.

All mutant forms of AfAMA were folded and gave CD spectra with pronounced α-helical minima at 222 nm and 209 nm similar to wild type AfAMA. The sigmoidal shape of the melting curves indicates that the unfolding process in all mutants was cooperative, and the melting temperatures of AfAMA mutants showed little variation from wild-type protein: 87±0.5°C for AfAMA Y19A, 86±1°C for AfAMA L21A and 87±1°C for AfAMA Y19A;L21A. AfAMA dimeric mutants did not show any significant chaperone activity when assayed with luciferase or citrate synthase as substrates (data not shown). Another striking result is that all dimeric mutants, defective in the GYPL motif in the N-domain, had ATPase activity similar to AfAMA-ΔN, which is dimer by itself (Fig. 4.41).
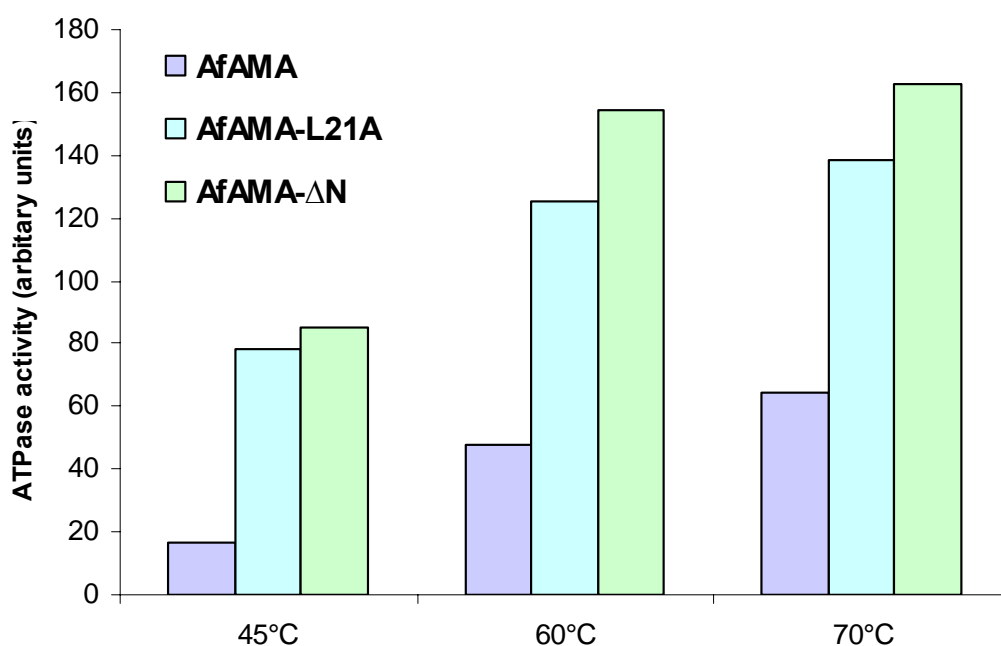
Figure 4.41 Comparison of the ATPase activity of AfAMA, AfAMA-ΔN and AfAMA-L21A measured at three different temperatures.

4.2.6 Chimeras of AMA and VAT-Nc

VAT-Nc domain of VAT protein from *T. acidophilum* (gi 6435755, residues 93-174) was cloned into the pET30b expression vector (Novagen) with *Nde*I and *Hind*III restriction sites. The construct was generated with a C-terminal 6xHis tag to facilitate purification. Proteins were expressed in *E. coli* C41 (DE3) at 37°C after induction with 1mM IPTG at OD_{600}~0.6. After purification using combined affinity and size exclusion chromatography, the protein was assayed for its possible chaperone activity. No activity was observed in assays with different protein substrates. Measurement of the native size on a calibrated size-exclusion column (Superose 12) indicated monomeric state of the protein with a molecular weight of approximately 10 kD. From these results and results obtained in Chapter 4.2.5, it became important to completely understand the role of the GYPL motif and the loop where this motif is settled. Based on the alignment of AMA-N domains with VAT-Nc (Fig. 4.39) we designed a couple of chimeric loop constructs. Loops following the first β-strand were either partly or totally interchanged between the β-clam domains of AfAMA-N and

VAT-Nc (Fig. 4.42). Constructs were generated using long forward primers, which introduced the loop changes and an N-terminal 6xHis-tag in the sequence to expedite protein purification. All constructs were expressed as soluble proteins in *E. coli* C41 (DE3) cells at 37°C after induction with 1mM IPTG at OD_{600}~0.6. Proteins were purified over Ni-NTA affinity matrix followed by gel filtration (Superdex G-75). Constructs were analyzed for their oligomeric state (Fig. 4.42) using calibrated Superose 6 and 12 columns.

```
                                                          oligomeric state

  >AfAMA-N
  MAKRETAELRYLIVRPLGYPLKASYHEYPQ-VDNPKV...                      6
  >AfAMA-N short
  MAKRETAELRYLIVRPLGYPLKASYH-----GEGPKV...                      6


  >VAT-AMA loop hybrid 1
  VRTEIAKKVTLAPIIRK----DQRLKFGEYPQVDNEG...                      1
  >VAT-AMA loop hybrid 2
  VRTEIAKKVTLAPIIRPLGYPLKASYHEYPQVDNPIE...                      3
  >TaVAT-Nc
  VRTEIAKKVTLAPIIRK----DQRLKF--GEGIEEYV...                      1
```

Figure 4.42 Oligomeric states of native β-clam domains (AfAMA-N and TaVAT-Nc) and the constructed chimeric proteins (AfAMA-N short and VAT-AMA loop hybrid 1(2)). The GYPL motif is colored in green; loops from AfAMA and VAT-Nc are designated with blue and red color, respectively.

Those chimeric β-clam domains that contained the GYPL motif had higher tendency to oligomerize (Fig. 4.42). The AfAMA-N construct that contained the GYPL motif but a shorter loop (AfAMA-N short), resembling by length the loop of VAT-Nc protein was a labile hexamer in comparison to the wild type AfAMA-N. The stability of the protein was drastically affected, which made it impossible for use in heat aggregation assays. It is important to mention again that deletion of the GYPL motif (AMA-N-ΔGYPL) in the AfAMA-N resulted in an insoluble product that we were not able to refold. Proteins were tested for possible chaperone activity in heat aggregation assays using citrate synthase as protein substrate. Chaperone activity of the chimera

proteins and wild type domains was compared using equimolar ratios of proteins and citrate synthase in the assay (Fig. 4.43). The trimeric VAT-AMA loop hybrid 2 was the only construct, besides wild type AfAMA-N protein, which was able to prevent heat aggregation of citrate synthase to some extent.
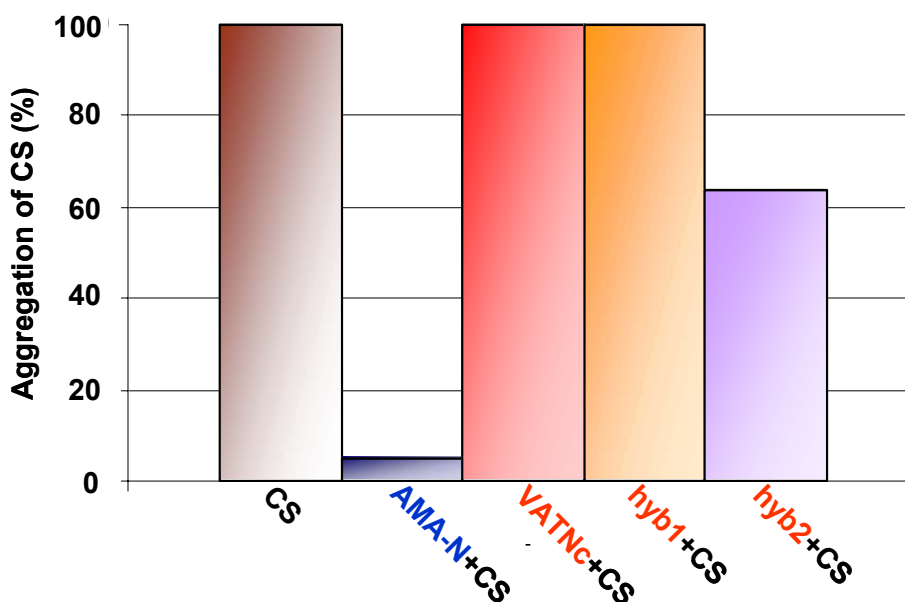

Figure 4.43 Citrate synthase heat aggregation assay in presence of AMA-N, VAT-Nc or two hybrids (hyb 1 and 2); each bar represents relative aggregation of CS after 40 minutes of heating at 50 °C

4.2.7 Structure determination of AMA proteins

Purified AfAMA, MjAMA and MjAMA-N proteins were set up for crystallization using the hanging or sitting drop method and crystal screens from Hampton Research and Sigma. First screens for AfAMA and MjAMA proteins in 20 mM MOPS, pH 7.25, 150 mM NaCl did not result in crystallization due to later observed uncontrolled aggregation in this buffer. Addition of 5% glycerol and presence of ADP or non-hydrolyzable ATP analogs in the sample buffer resulted in stable proteins that crystallized under several conditions (Fig. 4.44). Measurement of diffraction data and structure determination is an ongoing project. MjAMA-N crystals were obtained by the hanging drop method in the Index screen of Hampton research (0.05 HEPES, pH 7.5, 0.2 M KCl, 35% v/v Pentaerythritol Propoxylate (5/4 PO/OH) with different additives (Fig. 4.45).
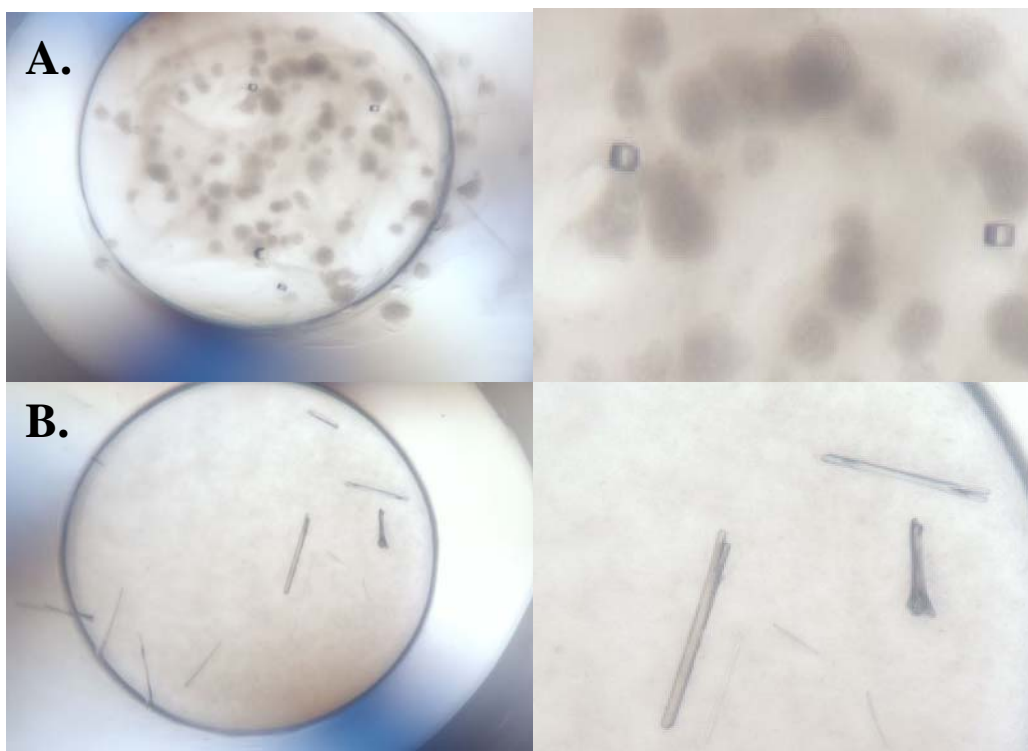
Figure 4.44 (A.) Crystals of the AfAMA with ADP (c = 10mg/ml, 0.1 M TRIS/HCl, pH 8.5, CaCl$_2$x2H$_2$O, precipitant 30% w/v PEG 4000); **(B.)** Crystals of AfAMA with ATPγS (c = 10mg/ml, 0.1 M TRIS/HCl, pH 7.0, 0.2 M MgCl$_2$x6H$_2$O, precipitant 2.5 M NaCl)

Diffraction data at a resolution of 2.8 Å was collected, and models for molecular replacement were generated based on the similarity of the AMA-N domain with VAT-Nc structure. We expect to solve the molecular structure of this domain in the neer future.
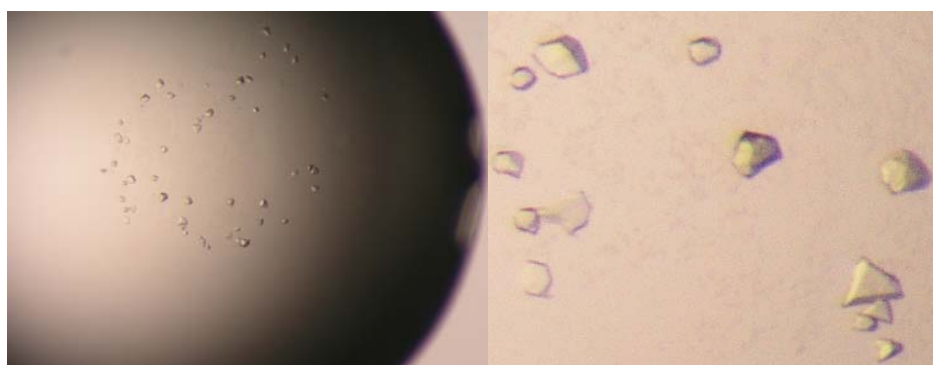

Figure 4.45 (A.) Crystals of the MjAMA-N domain (c = 9mg/ml0.05 HEPES, pH 7.5, 0.2 M KCl, 35% v/v Pentaerythritol Propoxylate (5/4 PO/OH)

Modeling of the AfAMA-N domain based on EM data and its similarity to VAT-Nc was done by Dr. Beate Rockel from the MPI for Biochemistry. The model was made by building a hexameric EM model of the beta-clam domain of VAT-N (residues 95-174) which was low-pass filtered to 2.0 nm using a pixel size of 2.65 Å. The hexameric model of VAT-Nc was used to build a surface and ribbon model of the AMA-N β-clam which was fitted into the refined EM data (Fig. 4.46). Based on data obtained from the GYPL-motif single point or deletion mutants, loops containing the motif were placed in the central pore of the hexamer. This orientation of the loops favours the proposed role of GYPL in oligomerization and/or binding of protein substrates.
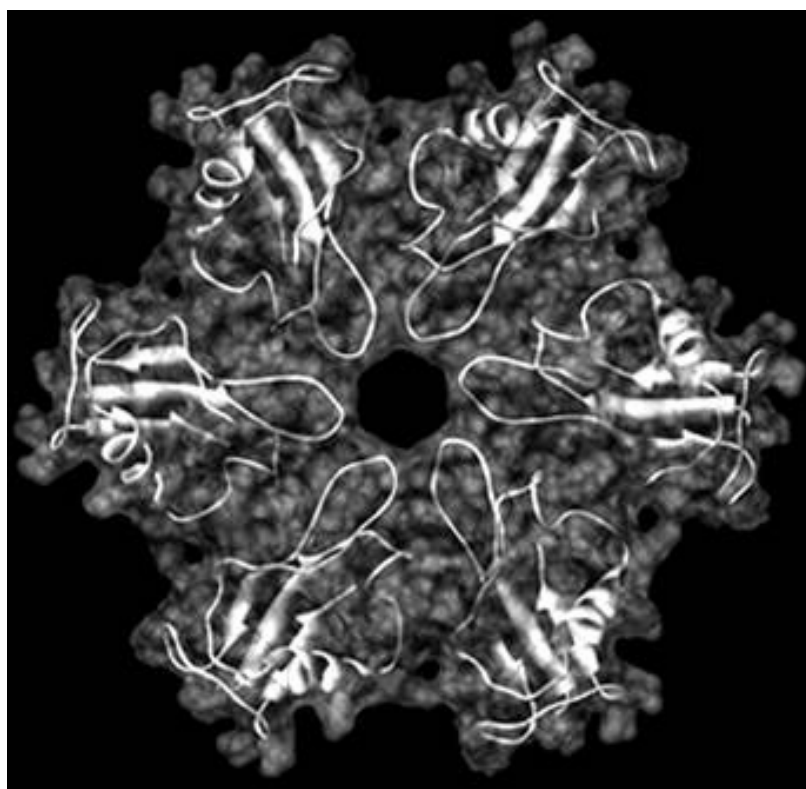

Figure 4.46 Ribbon and surface representation of the hexameric model of the AMA-N beta-clam domain

4.3 PAN-N and ARC-N domains

Clustering of the N-terminal domains of AAA proteins indicated a well defined group of PAN-N domains, which additionally contained N-domains of ARC proteins (Appendix Fig.1). Similarity between these sequence was not only significant in the coiled-coil segments (Wolf et al., 1998) but also in the β-strand rich interdomains (Frickey and Lupas, 2004). Since the structural data on ARC and PAN proteins were limited to the level of the electron microscopy methods (Wolf et al., 1998; Wilson et al., 2000; Zhang et al., 2004) we have tried to further characterize those proteins by crystallography or NMR spectroscopy. The comparative analysis of the chaperone function of the N-terminal domains of PAN and ARC was based on the knowledge that PAN and its eukaryotic homologs, subunits of the 19 S proteasome, can prevent aggregation of different protein substrates (Braun et al., 1999; Strickland et al., 2000; Benaroudj and Goldberg, 2000). Structural and functional characterization of the Ph1500 protein from *P. horikoshii* (gi 3257925) was done due to the fact that this protein was predicted to contain domains found in the N-termini of four different clades of AAA proteins (CDC48/p97, ARC, PAN and AMA). N-terminal part of the protein (residues 1-77) is predicted to assume β-clam fold found in the CDC48/p97 and AMA group, while C-terminus (78-148) shares structural similarity with β-strand rich domain of PAN and ARC proteins. Therefore, despite not being an AAA protein itself, Ph1500 is making an important link between the clades of AAA family.

4.3.1 Crystal structure of ARC-Nc domain

Constructs of ARC-N (residues 1–227, gi 3790601), ARC-Nn (1-77) and ARC-Nc (78-227) in pT7-7 vector (Zhang et al, 2004) were a generous gift from Dr. Peter Zwickl from MPI for Biochemistry in Martinsried. All constructs were generated with C-terminal 6xHis-tag to facilitate purification. ARC proteins were expressed in *E. coli*

BL21 (DE3) at 37°C after induction with 1mM IPTG at $OD_{600} \sim 1.0$ and purified by a combination of Ni-NTA affinity chromatography and gel filtration. Purified proteins were dialyzed overnight against 20mM MOPS, pH 7.25, with 120 mM NaCl buffer which was used for crystallization screens. Commercially available ready made crystal screens from Hamptone Research and Sigma were used for protein crystallization. ARC-N and ARC-Nc were set up for crystallization using the hanging drop method by mixing a 1:1 solution of protein and precipitant. ARC-Nc crystals of cubic shape were obtained using 30% v/v PEG monomethyl ether 550 as a precipitant, in 0.1 M Bis-TRIS, pH 6.5 buffer with 0.05 M $CaCl_2$ (Fig. 4.47).
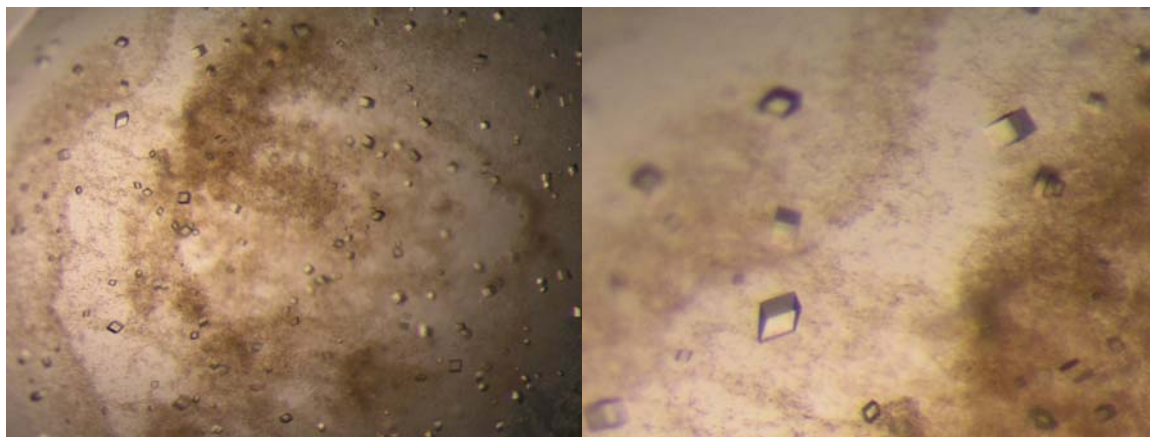

Figure 4.47 Crystals of the ARC-Nc (c = 3.6 mg/ml, 0.1 M Bis-TRIS, pH 6.5, 0.05 M $CaCl_2 x2H_2O$, precipitant 30 % v/v PEG monomethyl ether 550)

Crystal structure of the ARC-Nc domain was solved by Dr. Kornelius Zeth (from a dataset of up to 1.6 Å resolution. The structure reveals a hexameric complex of two almost identical OB-barrel folds (**o**ligosaccharide **b**inding) in each protomer (Fig. 4.48). Duplication of the OB folds is not obvious on the sequence level because of the 20 residue insertion in the loop following the helix of the second OB fold. This loop was also not seen in crystal structure, which is probably due to its flexibility (Fig. 4.49).
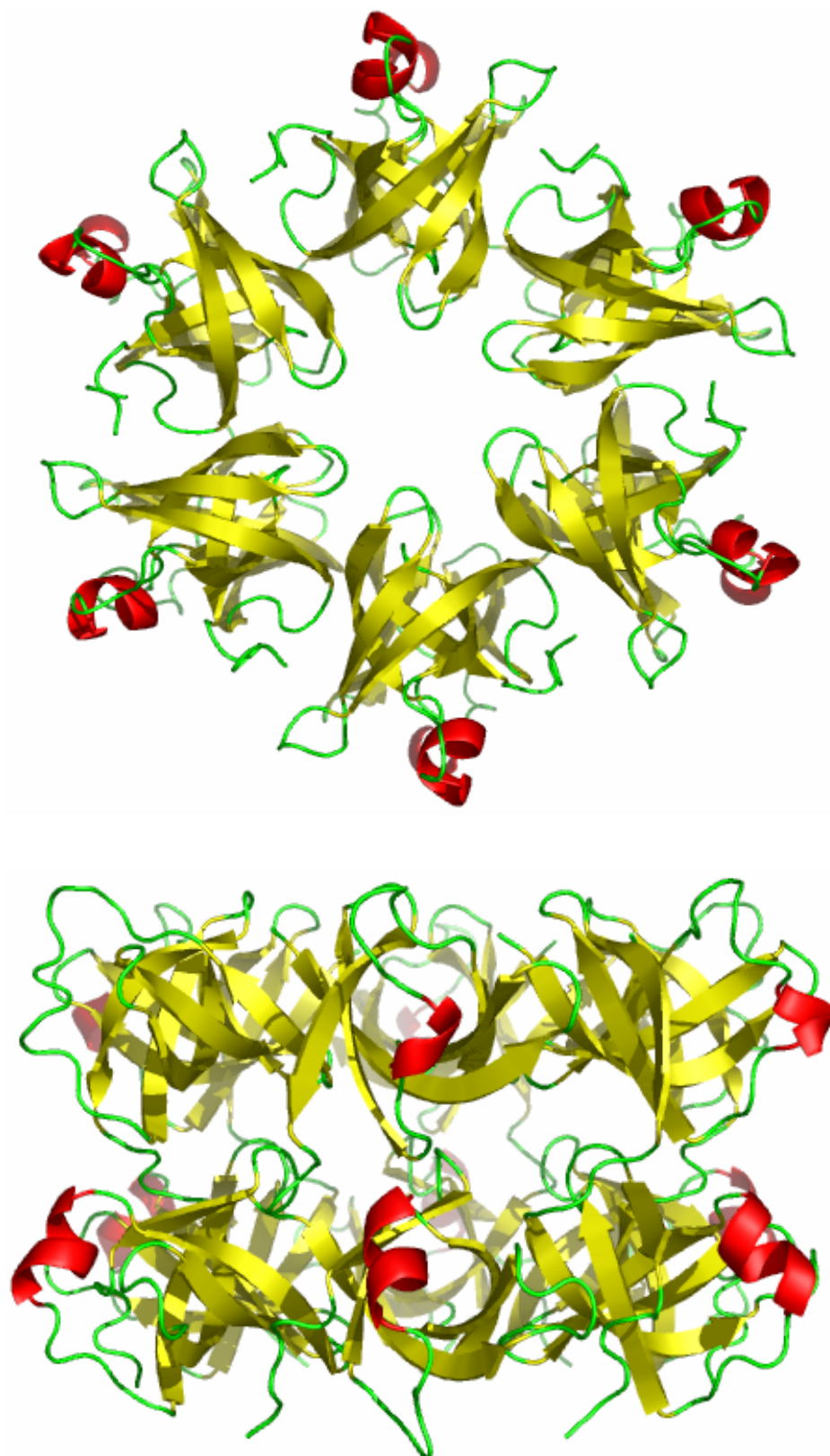
Figure 4.48 Crystal structure of the ARC-Nc (residues 78-227) hexamer complex in top and side view
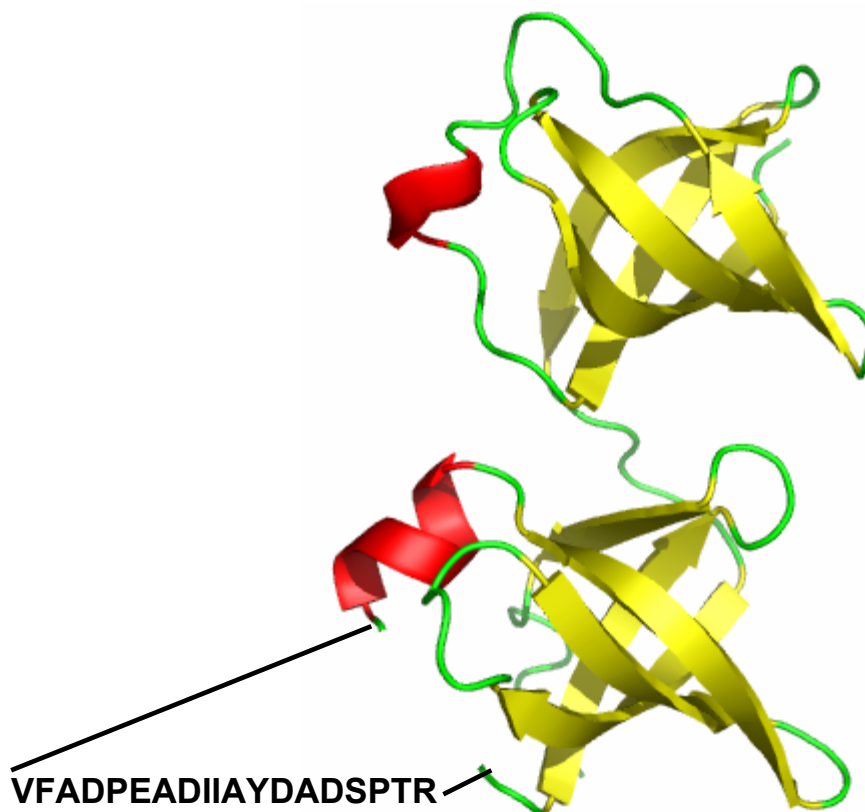
VFADPEADIIAYDADSPTR

Figure 4.49 Structure of ARC-Nc protomer showing position and sequence of the missing loop

4.3.2 Expression and characterization of PAN-N domains

DNA sequences encoding AfPAN-N (1-134) and its subdomains AfPAN-Cc (Cc-coiled coil, residues 1-60) and AfPAN-Ob (OB fold, residues 61-134) respectively, were amplified by PCR from genomic DNA of *A. fulgidus* and cloned into the pET30b vector (Novagen). All three constructs contained an N-terminal 6xHis tag to facilitate purification. For expression in the *E. coli* C41 strain, cells were grown in LB medium at 37°C, induced at OD_{600}~0.6 with 1 mM IPTG and harvested after 4h. AfPAN-Cc and AfPAN-Ob were purified by a combination of Ni-NTA affinity chromatography and gel filtration on a Sephadex G-75 (GE Healthcare) column. Purity of the proteins (>95%) was checked using SDS-PAGE. The majority of AfPAN-N was found in inclusion bodies. The protein was solubilized in buffer A1 containing 8M urea, and purified by Ni-NTA affinity chromatography under

denaturing conditions. 95% of the protein was refolded by overnight dialysis against 50 mM Na-Phosphate, pH 7.4, 150 mM NaCl, 1mM DTT and 8% glycerol. To obtain information about the molecular mass and oligomeric state of all constructs, samples were analyzed on analytical Superose 6 and 12 columns (GE Healthcare), calibrated with gel filtration standards (Sigma) and additional standard proteins. While AfPAN-Cc and AfPAN-Ob were behaving like dimers, the AfPAN-N domain formed soluble oligomers with high molecular weights in the range of 1 MDa. Electron microscopy of the protein sample showed round shaped complexes of uniform size (Fig. 4.50)
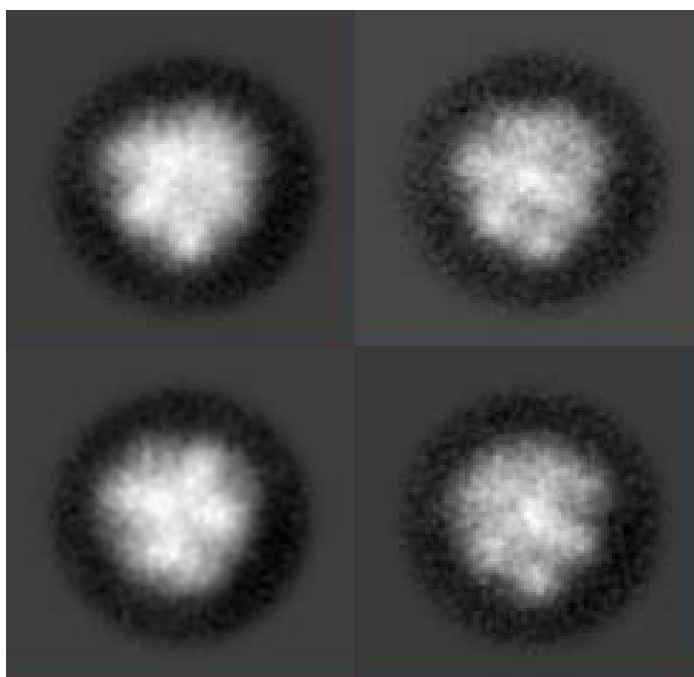
Figure 4.50 Class averages of the uranyl-acetate stained AfPAN-N particles

Instability of the AfPAN-Ob subdomain and the size of AfPAN-N particles rendered them unsuitable for structure determination by NMR spectroscopy. The purified N-terminal domain of *M. janaschii* PAN protein behaved like a dimer but also showed instability. Expression of PAN-N domain with N-terminal thioredoxin fusion gave soluble protein. After purification on Ni-NTA affinity chromatography and gel filtration the protein behaved like a dimer. The fusion protein was dialyzed against 20 mM MOPS, pH 7.2, 150mM NaCl buffer and set up for crystallization, but no crystals were obtained. The complete PAN protein (PAN-N+AAA domain) from *A. fulgidus*

was cloned and purified for the structural studies. Crystallization of PAN is still an on-going project in the moment.

4.3.3 Chaperone activity of ARC and PAN N-domains

Based on prior knowledge that full PAN protein from *M. janaschii* can prevent heat-induced aggregation of protein substrates (Benaroudj and Goldberg, 2000) N-terminal domains (and sub-domains) of PAN and ARC were assayed for possible activity against different protein substrates. Chaperone activity was restricted to fully expressed N-domains (Fig. 4.51 and 4.52), while sub-domains alone did not show any activity (data not shown).
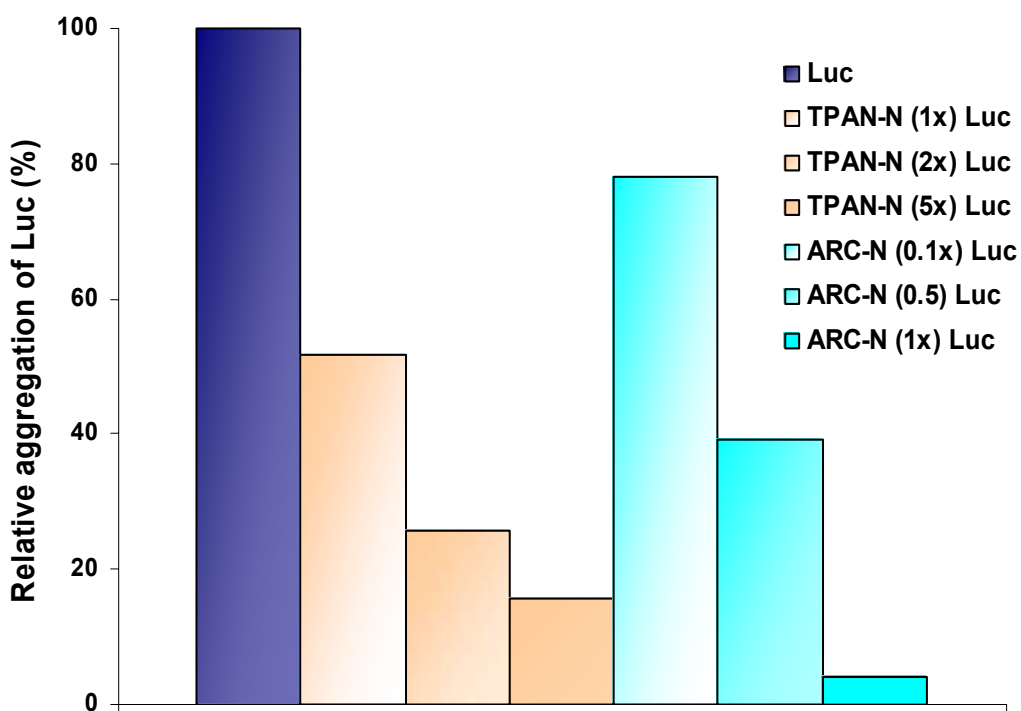

Figure 4.51 Luciferase heat aggregation assay (43°C) in the presence of different concentration of Thioredoxin-PAN-N (TPAN-N) and ARC-N. Numbers in brackets represent the molar ratio of the assayed proteins towards citrate synthase. Note: lower % of the relative aggregation means higher chaperone activity.
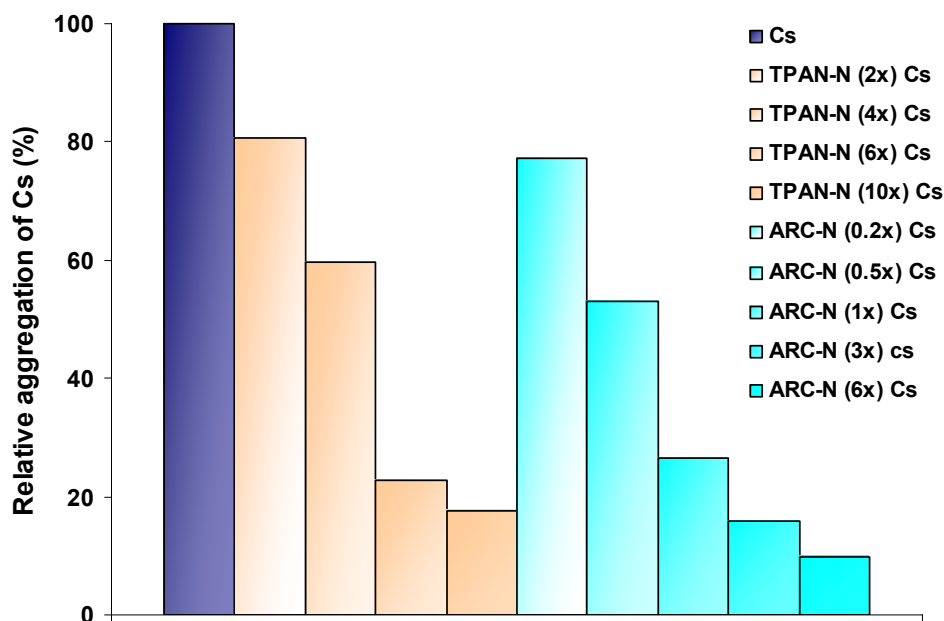
Figure 4.52 Citrate synthase heat aggregation assay (50°C) in the presence of different concentrations of Thioredoxin-PAN-N (TPAN-N) and ARC-N. Numbers in brackets represent the molar ratio of the assayed proteins towards citrate synthase.

Pull-down assays using Ni-NTA beads showed that citrate synthase (Fig. 4.53) and luciferase (not shown) can interact with PAN-N and ARC-N only when they are unfolded, the control with a non-heated mixture resulted in binding of 6xHis-tagged protein only.
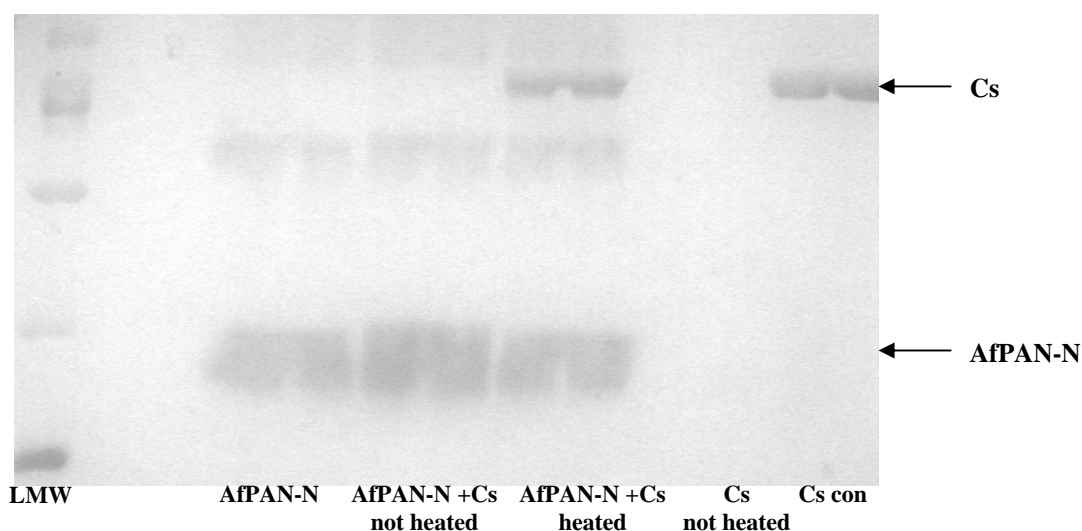
Figure 4.53 Ni-NTA pull-down after citrate synthase (Cs) heat aggregation assay in presence of AfPAN-N. Cs con – represents total input of citrate synthase in aggreg. assay. Cs not heated – is the control for unspecific binding of CS to Ni-NTA.

4.3.4 Chaperone activity of chimeric constructs

It was noticeable after the first assays that only full N-domains, containing both coiled-coil and OB fold(s) were active in assays. Sub-domains by themselves, coiled coil or OB fold(s), could not prevent aggregation of any tested substrate, which illustrates the importance of the structural arrangement of these domains in one polypeptide chain. To test if the unique combination of the coiled-coil and OB folds in ARC-N and PAN-N makes these proteins active ARCPAN or PANARC chimeras were constructed. In similar fashion 'frankenstein' chimeras were constructed where either one or both sub-domains were exchanged with a structurally but not functionally similar domain. The coiled-coil segment of the GCN4 transcription factor from *S. cerevisiae* was used to substitute ARC-N and PAN-N coiled-coils. OB folds of the translation initiation factor from M. janaschii (2EIF) and RNA-binding cold shock protein from E. coli (1MJC) were found using 3D protein structure comparison and alignment (http://cl.sdsc.edu/; Shyndyalov & Bourne, 1998) with OB folds of ARC-N (Fig. 4.54). By joining GCN4 coiled-coil with the mentioned OB folds we have constructed proteins that beside structural architecture had nothing in common with the original N-domains of PAN and ARC.

1MJC	**ARC-N**	**2EIF**

Figure 4.54 Comaparison of the OB-folds from ARC-N, *E. coli* cold shock protein (1MJC) and translation initiation factor from *M. janaschii* (2EIF). The two structures were used to construct 'frankenstein' proteins.

Influence of the length of the coiled-coil segment on the activity of the N-domains was elucidated through a series of PAN-N constructs with shorter coiled-coil sub-domains (Table 4.1).

Construct	Origin
ARCPAN	cc ARC + OB AfPAN
PANARC	cc AfPAN + OBs ARC
GARC	cc GCN4 + OBs ARC
GPAN	cc GCN4 + OB AfPAN
AECS	cc ARC + OB EcCS
ATIF	cc ARC + OB MjTIF
GECS	cc GCN4 + OB EcCS
PAN-N 3hep	AfPAN-OB with cc of 3 heptades
PAN-N 2hep	AfPAN-OB with cc of 2 heptades
PAN-N 1hep	AfPAN-OB with cc of 1 heptades
ARC-N short	ARC-N without second OB fold

Table 4.1 Chimera (blue text), 'frankenstein' (orange) and shortened (green) ARC/PAN-N-like proteins; cc - coiled-coil, OB - OB fold; EcCS - *E. coli* cold shock protein; MjTIF - *M. janaschii* translation initiation factor.

All proteins had an additional N-terminal 6xHis-tag to facilitate purification and were expressed from pET30b vector. After expression in the *E.coli* C41 strain, soluble proteins were purified by a combination of Ni-NTA affinity chromatography and gel filtration. The content of secondary structure of all proteins was tested by CD spectroscopy. Stability of proteins was tested by proteinase-K treatment and thermal denaturation. The ARC-N short construct was unstable at temperatures above 42°C, all other proteins had melting temperatures above 55°C. The activity of the chimeras, 'frankenstein' and shortened proteins was tested in heat aggregation assays with luciferase and citrate synthase and compared with the activity of native ARC/PAN-N domains (Fig. 4.55).
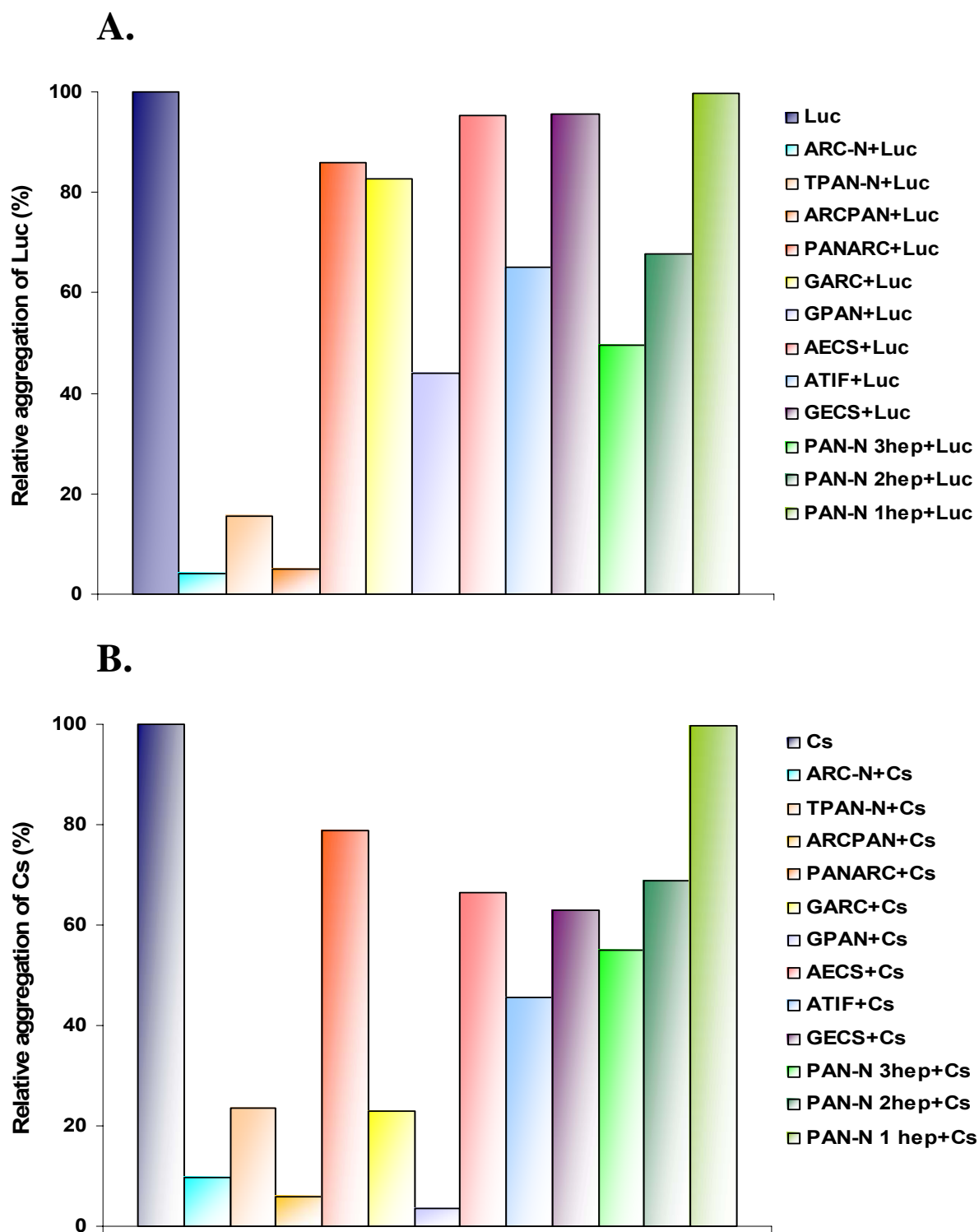
Figure 4.55 Heat aggregation assays of (A) luciferase (Luc) and (B) citrate synthase (Cs) in the presence of different proteins. Assays were performed with a five-fold excess of the protein of interest. Note: lower % of the relative aggregation means higher chaperone activity; for protein naming see the table 4.1.

Most of the chimeras and 'frankenstein' proteins were able to prevent heat aggregation of both substrates with preferences towards citrate synthase. Shortening of the coiled-coil region in the PAN-N construct (PAN-N 3-1hep) reduced activity of those proteins, where the PAN-N 1hep, with coiled-coil of only one heptade did not show any activity at all.

4.3.5 Chaperone activity of PP-linker mutants

Alignment of N-domains of ARC and PAN proteins from different organisms showed that besides their conservation of domain architecture, coiled-coil followed by OB fold, the connection between the two sub-domains is also highly conserved (Fig. 4.56). Two prolin residues between the two sub-domains, that we named PP-linker, are almost invariant in all sequences.

PAN-N

```
      hhhhhhhhhhh    ssssssss   ssssss   sssss
>gi|3122632|Archaeoglobus fulgidus
…EREVRRLRSEVERLRSPPLLVGVVSDILEDGRVVVKSSTGPKFVVNTSQ…
>gi|3256587|Pyrococcus horikoshii OT3
…ERELSRLRSEMSRLRQPPAFAGTVIEVLDEDRAIVQNYNGPRFVVRIAP…
>gi|15669365|Methanocaldococcus jannaschii
…LKENEILRRELDRMRVPPLIVGTVVDKVGERKVVVKSSTGPSFLVNVS…
>gi|3122631|Methanothermobacter thermautotrophicus
…DREVKSLRGEIERFRTPPLVIATVTEVLDDHRVAVKSTTGPHFVINYSR…
>gi|25289940|Sulfolobus solfataricus
…RKELNYYKAEMEKMLSPPLIEAVVLDVLPDGRVLVRSSSGPNLVVNIAS…
```

ARC-N

```
      hhhhhhhhhhh    ssssssss   sssss   sssss
>gi|3790601| Rhodococcus erythropolis
…RQQLIALREEVDRLGQPPSGYGVLLSVHEDKTVDVFTSGRKMRLTCSPN…
>gi|54040417|Mycobacterium bovis
…RQQLLALREEVDRLGQPPSGYGVLLATHDDDTVDVFTSGRKMRLTCSPN
…
>gi|467000|Mycobacterium leprae
…RQQLLALREEVDRLGQPPSGYGVLLAAHDDETVDVFTSGRKMRLTCSPN
…
>gi|29833219|Streptomyces avermitilis
…RDQIVALKEEVDRLAQPPAGFGVFLTANEDGTADIFTGGRKLRVNVSPS…
>gi|19552711|Corynebacterium glutamicum
```

Figure 4.56 Alignment of connector region in PAN-N and ARC-N sequences. Green and red residues represent parts of coiled-coil and OB fold sub-domains. The PP-linker is shown in blue. The consensus secondary structure for each group is listed above the sequences (h/helix, s/strand).

Based on the structure of ARC-N OB folds, the second prolin residue is embedded between two protomers. Thus it might be crucial for the correct projection of the coiled-coil part from the core of the N-domain made up of OB folds. This implied an importance of these prolines in the chaperone activity of these proteins. To check if the two residues are important, we constructed a series of mutants with ARC-N and GPAN shown in Figure 4.57.

ARC-N

…LREEVDRLGQ**PP**SGYGVLLSVHE…

ARC-N(PA)

…LREEVDRLGQ**P**ASGYGVLLSVHE…

ARC-N(PW)

…LREEVDRLGQ**P**WSGYGVLLSVHE…

ARC-N(AA)

…LREEVDRLGQAASGYGVLLSVHE…

ARC-N(AP)

…LREEVDRLGQA**P**SGYGVLLSVHE…

GPAN

...KNYHLENEVARLRS**PP**LLVGVVSD…

GPAN(PA)

...KNYHLENEVARLRS**P**ALLVGVVSD…

GPAN(PW)

...KNYHLENEVARLRS**P**WLLVGVVSD…

GPAN(AA)

...KNYHLENEVARLRSAALLVGVVSD…

GPAN(AP)

...KNYHLENEVARLRSA**P**LLVGVVSD…

Figure 4.57 Mutants of ARC-N and GPAN in PP-linker region. Residues belonging to coiled coil and OB fold are marked green and red, respectively. The PP-linker is marked in blue with introduced mutations in black.

Mutants were expressed and purified with the same protocol used for the wild type proteins and tested for their chaperone activity in heat aggregation assays with luciferase and citrate synthase (Figure 4.58). Results were similar for the ARC-N and GPAN mutants, with the single alanine replacements preventing 60% and 80% of heat aggregation and the double replacement showing no chaperone activity. An exception was the PW mutant, as GPAN (PW) showed even greater chaperone activity then the unmutated chimera. ARC-N (PW) did not show greater activity from wild type protein, but reduction in the activity was much lower in comparison to the other mutants.
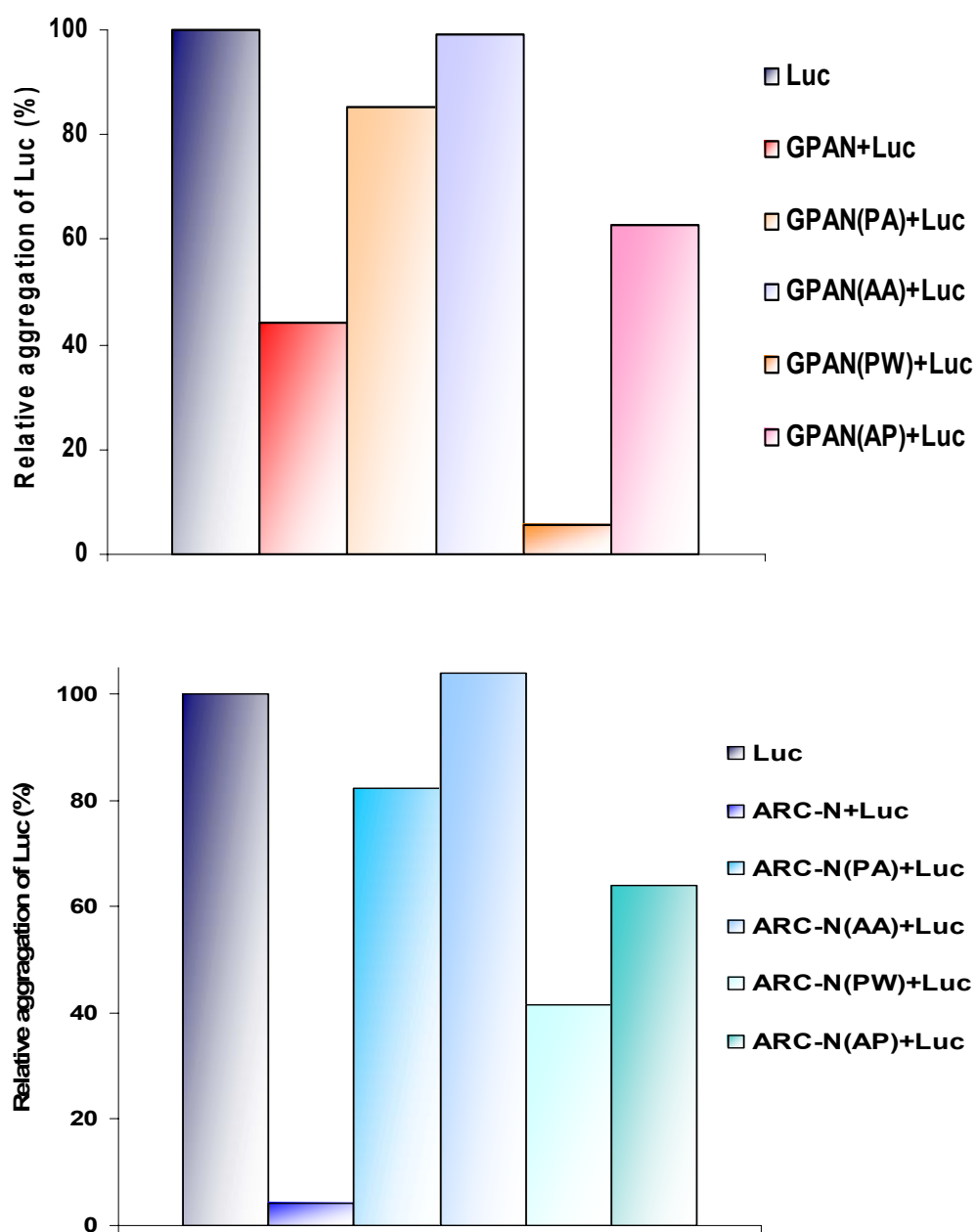
Figure 4.58 Heat aggregation assays of luciferase (Luc) in the presence of a different PP-linker mutants. Assays were performed with a five-fold excess of the protein of interest.

4.3.6 Expression and characterization of Ph1500 and its domains

The sequence of the Ph1500 protein from *P. horikoshii* (gi 3257925) was found from a PSI-BLAST search using VAT-Nc as the query sequence. Besides a β-clam domain (residues 1-77), the protein contains an additional second domain (residues 78-148), which was predicted to adopt the OB-fold. The OB-fold is usually designated as a structure that binds nucleic acids, but our studies on ARC and PAN proteins from the AAA family revealed the importance of this domain in the binding of protein substrates (Chapter 4.3). Therefore, despite not being a AAA protein itself, Ph1500 contains domains found in the N-termini of four different clades of AAA proteins (CDC48/p97, ARC, PAN and AMA), making an important link between these clades. An additional BLAST search using the sequence of Ph1500 protein as a query spotted 3 homologous sequences, two from *Pyrococci* (*P. abyssi, P. furiosus*) and one from *T. kodakarensis*. Analysis of gene loci conservation with "The SEED" server indicated gene coupling of the Ph1500-like proteins with a gene that is annotated as an endonuclease III (Fig. 4.59), an enzyme which makes excision repair of mismatched G-T pairs from damaged DNA molecules.
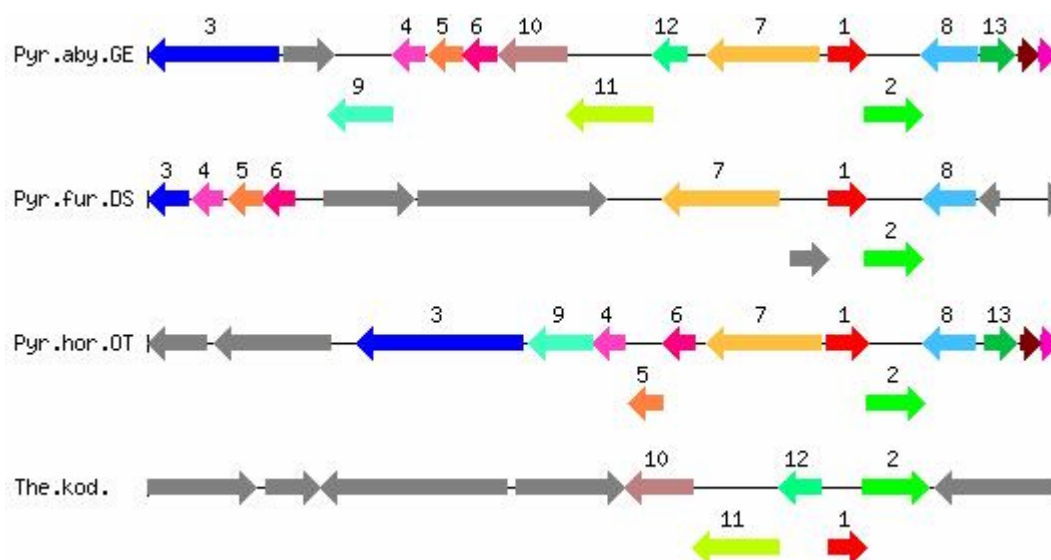

Figure 4.59 Analysis of the Ph1500 gene locus using The SEED. Ph1500-like genes are labeled with red color and number 1; endonuclease III genes are labeled with green color and number 2. The genes were found only in the genomes of *P. abyssi, P. furiosus*, *P. horikoshii* and *T. kodakarensis*.

The DNA sequence of the Ph1500 protein and its domains Ph1500N (residues 1-77) and Ph1500C (residues 78-148) were amplified from genomic DNA of *P. horikoshii*. Constructs were cloned into the pET-30b expression vector (Novagen) with *Nde*I and *Hind*III restriction sites for Ph1500 and Ph1500N, and with *Nde*I and *Xho*I for Ph1500C. All proteins were generated with an N-terminal His-tag to facilitate purification. Proteins were expressed in *E. coli* C41 (DE3) at 37°C after induction with 1mM IPTG at OD_{600}~0.6. After purification, using a combination of Ni-NTA affinity and gel filtration chromatography, the proteins were dialyzed against the buffer for crystallization screens (10mM MOPS, pH 7.25, 150 mM NaCl). This buffer was also used for subsequent crystallization screens. Protein samples were run over a calibrated size exclusion column (Superose 12) to figure out their oligomerization state and molecular weight. Ph1500 and Ph1500C were eluted as oligomers with estimated molecular weights of 110 kD and 55 kD, respectively. These are close to the calculated molecular weights for hexameric complexes of these constructs: 102 kD for Ph1500 and 54 kD for Ph1500C. The Ph1500N domain eluted as a monomer of approximately 10 kD (9.5 kD calculated). Constructs were analyzed for possible interaction with unspecific DNA fragments, but no binding was observed. The endonuclease III from *P. horikoshii* was also cloned and expressed, and purified protein mixed with the Ph1500 or its N- and C-domains. The proteins were tested for co-migration on a calibrated gel-sizing column, indicating complex formation. First results showed an interaction of endonuclease III with either N-domain (Ph1500N) or full Ph1500 protein. However, these results need to be confirmed by NMR or pull-down assays. Ph1500 protein will be also assayed for interaction with single-stranded DNA, as well as with mismatch double-stranded DNA mimicking DNA damage.

4.3.7 Electron microscopy and NMR structure of Ph1500

Preliminary microscopy studies on Ph1500 (Fig. 4.60A) were done at low resolution at our in-house electron microscopy unit with kind help of Dr. Heinz Schwarz. Data collection of negatively stained Ph1500 protein particles (uranyl acetate) as well as selection and averaging of the obtained images was done by Dr. Beate Rockel from

the MPI for Biochemistry in Martinsried. 42 images were recorded on a CCD camera (24 μm pixel size) with a primary magnification of 45K and a post magnification of 1,9, resulting in a final pixel size of 2.8 Å. From the images, 10003 particles were selected and subjected to translatory alignment in order to center the particles properly. An eigenvector-analysis of the aligned data set produced class averages with clear six-fold symmetry (Fig 4.60 B). Rotational alignment using hexameric class-average revealed particles of two different ring sizes (Fig 4.60 C and D).
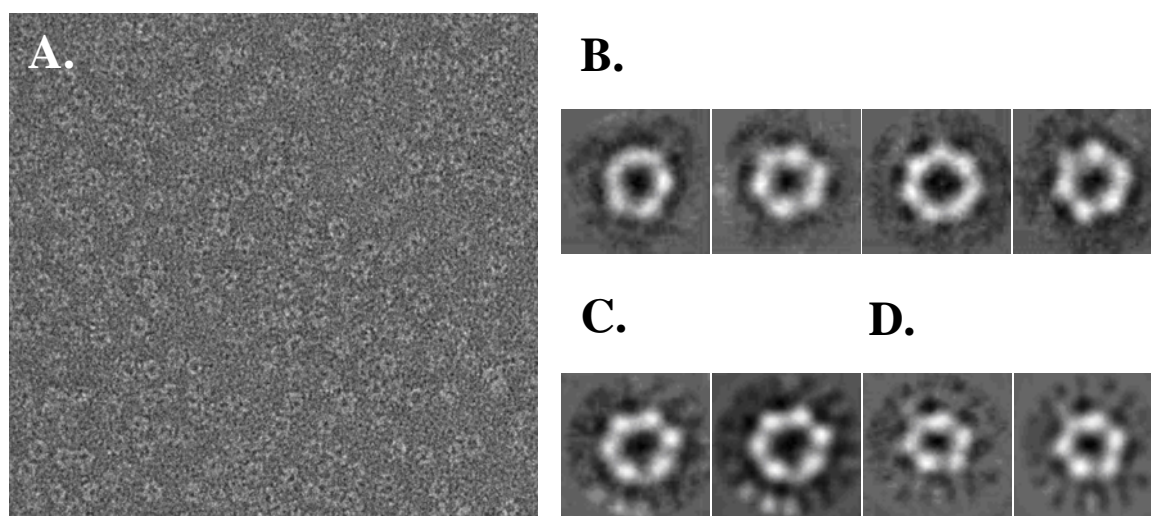
Figure 4.60(A.) Ph1500 particles negatively stained in uranyl acetate. **(B.)** Class averages after eigenvector-analysis **(C. and D.)** Class averages with different ring size after rotational alignment.

Since the crystallization screens did not provide crystals of Ph1500, we decided to solve the structure of this protein by NMR spectroscopy. The size of the Ph1500 hexamer (106 kD) is far beyond usual limits of NMR studies, but the fact that the domains of the protein can be expressed separately and exhibit high stability enabled us to do structure determination by NMR spectroscopy. The Ph1500N domain was expressed in minimal media supplemented with ^{15}N-labeled NH_4Cl as the sole nitrogen source, and ^{13}C-labeled glucose as the sole carbon source. The structure of the N-domain was determined by Ilka Varnay (PhD student at Technical University of Munich, Garching) and it resembles the β-clam fold of the VAT-Nc domain with noticeably shorter loops and prolonged secondary structure elements (Fig. 4.61).

Structure determination of the C-terminal domain (Ph1500C) is an on-going project. For determination of the hexameric structure of Ph1500C, we are using a triple labeling strategy. In addition to the usual single or double labeling (^{15}N, ^{13}C) we are deuterating the sample up to 90% to obtain clear ^{15}N-^{13}C coupling data. Selective unlabeling of branched amino-acids (Leu, Ile, and Val) is another method that we are currently using to obtain selectively H-H couplings for the side chains interacting between two neighboring Ph1500C domains. The structure of the complete protein will be calculated from similar experiments using existing data on individual domains.

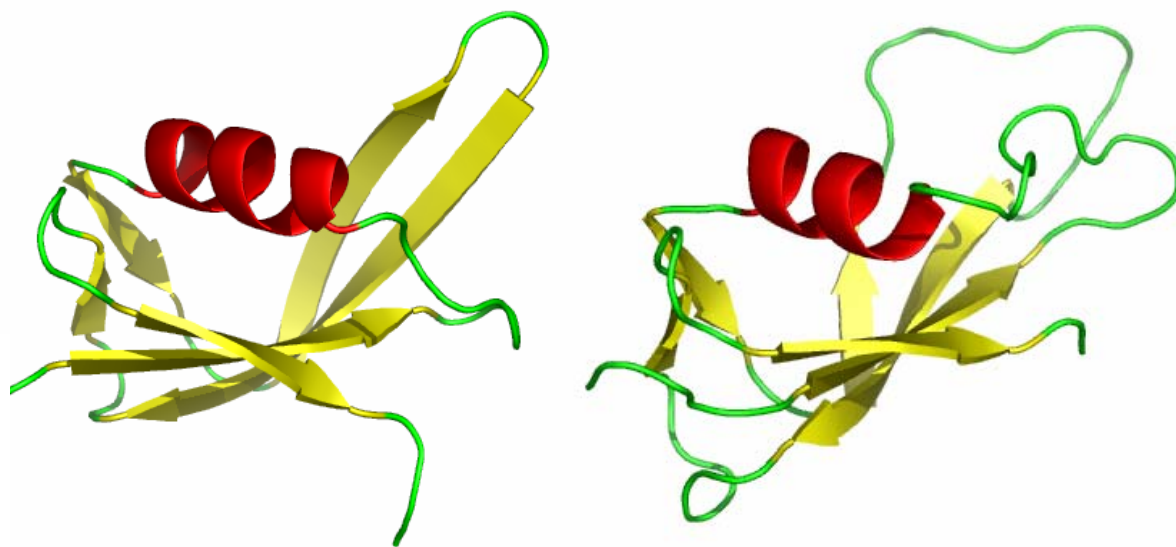**Figure 4.61** Comparison of the NMR structures of the Ph1500N (left) and VAT-Nc domain (right, 1CZ4 89-175). Domains are orientated in a way that similarities and differences between them can be easily noticed.

5. Discussion

5.1 Cradle-loop barrels

5.1.1 Double-psi barrel

Search for the experimental evidence that would sustain the proposed evolutionary path of the double-psi barrel fold (Coles et al, 1999) revealed an even more complex picture in the evolution of this fold. The chase for the simple 40 residue βαββ motif that would give rise to the members of the double-psi barrel fold through circular permutation we have started from the VAT-Nn sequence by reshaping it in a way to exemplify retrograde evolution (Fig. 5.1).

```
                ssssss              ssss hhhhh          ssssss
VAT-Nn Ta       mesnnGIILRVAEANSTDPG-MSRVRLDESSRRLLDAE-IGDVVEIEKVr
                KTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV
                                    ↑   divergence of halves

VAT-Nnc db      mvKTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKVgh
                KTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV
                                    ↑   duplication of ββαβ element

VAT-Nnn         mesnnGIILRVAEANSTDPG-MSRVRLDESSRRLLDAE-IGDVVEIEKVr
VAT-Nnc         mvKTVGRVYRARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV
                                    ↑   circular permutation

VAT-Nnn cp          mANSTDPG-MSRVRLDESSRRLLDAE-IGDVVEIEKVr--GIILRVA
VAT-Nnc cp          mARPEDEN-KGIVRIDSVMRNNCGAS-IGDKVKVRKV---KTVGRVYR
                ssssss              hhhhhhh     sssssss   sssssss
Bs AbrB-N       MKSTGIVRKVDE-----LGRVVIPIELRRTLGIA-EKDALEIYVDD—EKIILKKYkpnmt
```
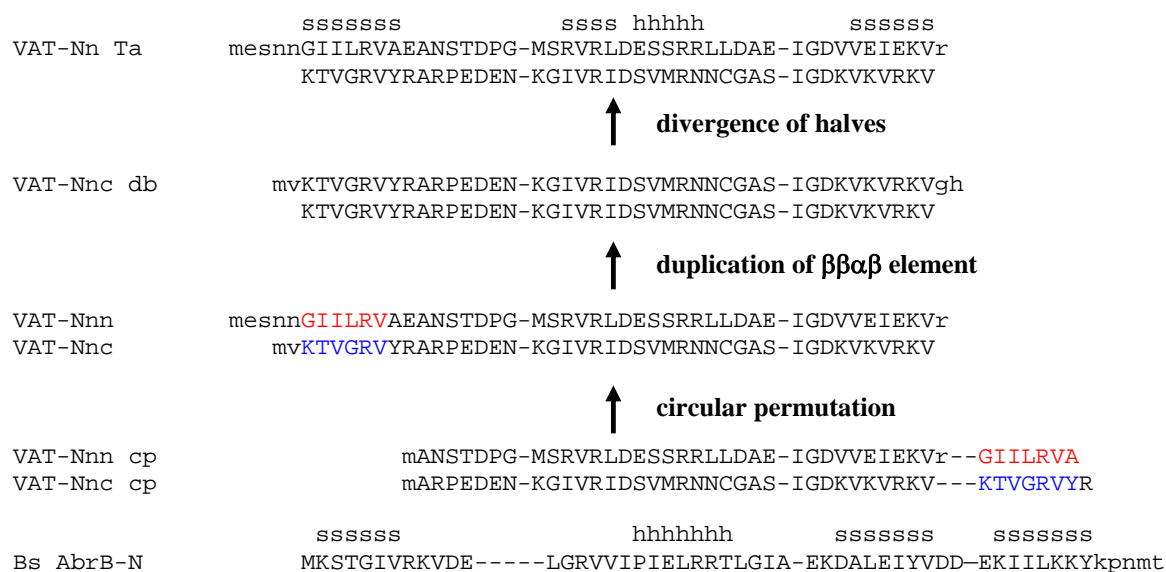
Fig 5.1 Evolution of the double-psi barrel fold made through constructs based on VAT-Nn sequence and similarity to AbrB-N. Postulated events are written next to each step. Experimental path was done in different direction. Circular permutation of the strands is shown in color for clarity. The secondary structure elements are indicated above the alignment; s=strand, h=helix.

The construction of a double-psi barrel sequence with 'equal' halves by doubling the sequence of either VAT-Nnn or VAT-Nnc in one polypeptide chain resulted in folded proteins. Moreover, NMR spectra of VAT-Nnc db (doubled) could be compared to the spectra of the wild type VAT-Nn sequence. Such result is not surprising taking into account that sequence identity between the two halves of VAT-Nn is 38% over 42 residues, which make up the core of the fold. Stability of VAT-Nnn db was much lower compared to the wild type VAT-Nn and the VAT-Nnc db construct and when the halves were expressed separately, without duplication (VAT-Nnc and VAT-Nnn), only VAT-Nnc was folded. It is interesting that similarity of folding properties of VAT-Nn and VAT-Nnc in making two distinguished species was further found at the functional level. Chaperone activity of the dimer and tetramer of VAT-Nn and VAT-Nnc, respectively, would lead to the conclusion that a structural complex of four ββαβ-repeats is needed for binding of protein substrate. Unfortunately we were not able to determine the structure of either the VAT-Nn dimer or the VAT-Nnc tetramer by NMR spectroscopy. It can be speculated that two double-psi barrels formed in these two cases would take a side by side orientation with their psi-loops orientated towards one side, where one of the barrels would take the position of the beta-clam domain (VAT-Nc) (Fig. 5.2 A). Another more favorable orientation would be loops-to-loops arrangement which would resemble fold and substrate binding site of aspartic proteases (Fig. 5.2 B). Such a structural arrangement would further support the postulated evolutionary path, in which two double-psi barrels will at the end give rise to complex fold of aspartic proteases (5.3).
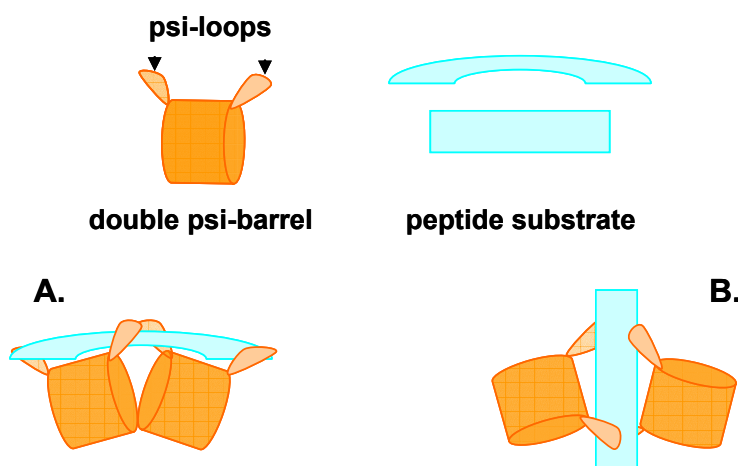

Figure 5.2 Two schematic models of VAT-Nn dimer binding substrate (**A**) side by side and (**B**) loops-to-loops.
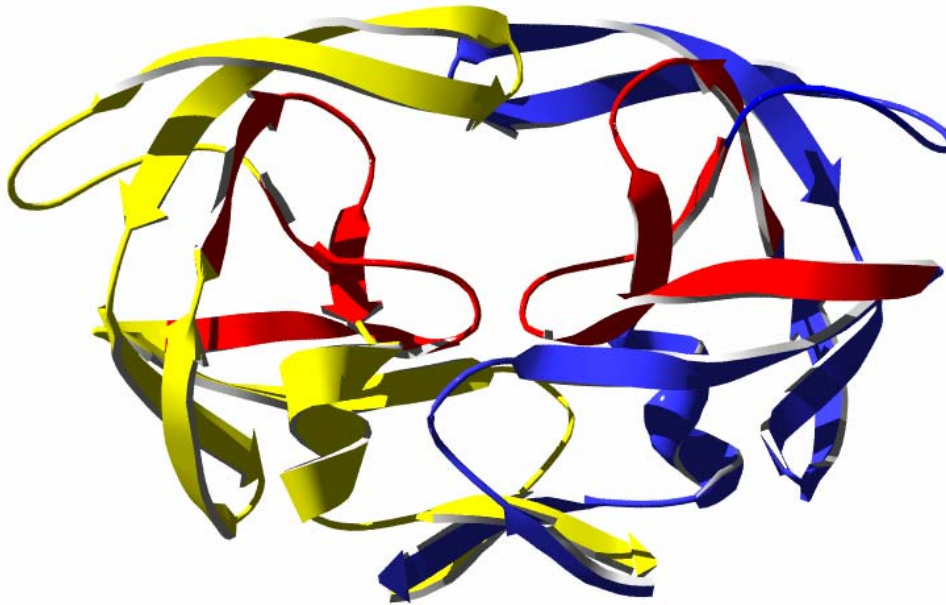
Figure 5.2 Crystal structure of HIV protease (1W5X). Two double-psi barrel protomers are shown in yellow and blue; psi-loops and β-hairpins forming the binding sites are colored red.

The similarity of VAT-Nn with AbrB-N and other proteins that were annotated to be transcriptional regulators (Mj0056, SpoVT, PemI) was further confirmed with DNA shift mobility assays in which only the monomeric VAT-Nn domain showed unspecific DNA binding. Competitive inhibition of the VAT-Nn chaperone activity where the DNA substrate was added to the heat aggregation assay with citrate synthase did not answer any question. First, because the VAT-Nn dimer did not show any DNA binding and second, the monomer of VAT-Nn did not exhibit any chaperone activity. Originally it was postulated that an 'ancestral' double-psi barrel had substrate binding sites between helices capping the barrel and loops following the first beta strands of each half of the barrel (Coles et al., 1999). Evolution of the side loops would result in another binding site that would be placed between two loops. Based on the conserved positions of the Arg-residues (Fig1.9) and the charge distribution in the surface model of VAT-Nn barrel (Fig. 1.8), it can be deduced that binding of the DNA substrate would be equally favorable at both proposed sites, "original" and "evolved". Fine mapping of the substrate binding in both cases using

NMR spectroscopy and determination of the VAT-Nn dimer structure would give more answers to possible binding scenarios.

5.1.2 AbrB –swapped hairpin barrel
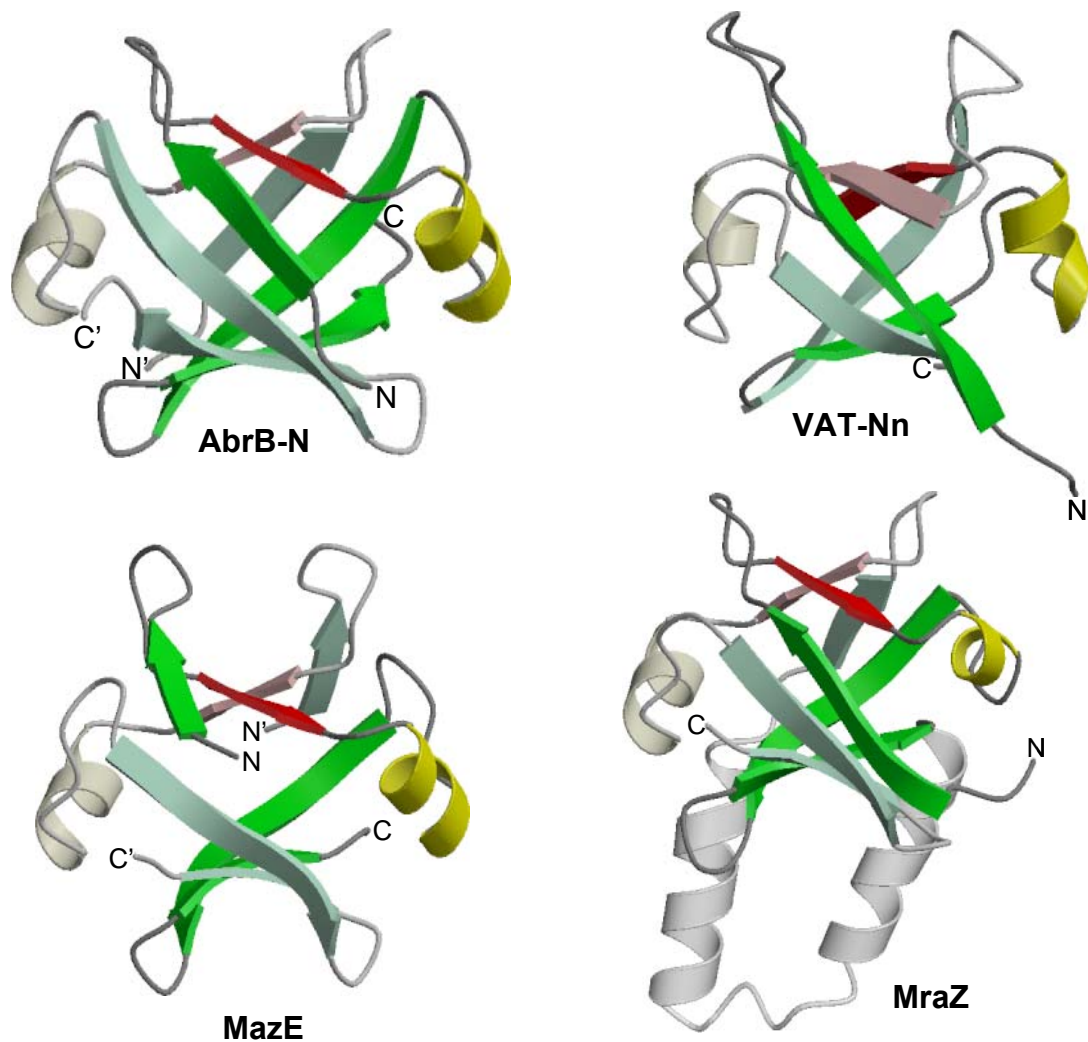
Figure 5.4 Structures of the swapped hairpin barrels (AbrB-N, MazE and MraZ) and related double-psi barrel (VAT-Nn). The secondary structure of one symmetrical half of each domain is shown in the bold color, with the β2 strand in red. In the structure of MraZ one helical hairpin is left out for better representation of the barrel structure.

The discrepancy of the AbrB-N structure from Cavanagh's group (1EKT; Vaughn et al., 2000) with the prediction that was made based on the sequence similarity with VAT-Nn (Coles et al., 1999) and moreover with the crystal structures of the homology-related MazE (1MVF; Loris et al., 2003) and MraZ protein (1N0G; Chen et

al., 2004) prompted us to redetermine the structure of AbrB-N. The newly solved structure of the dimeric AbrB-N (1YSF; Coles et al., 2005) was in agreement with the crystal structures of its homologous, thus forming an eight stranded swapped-hairpin barrel (Fig. 5.4). Each monomer of AbrB-N consists of four β-strands arranged in two hairpins which are interleaved in the dimer. Two folds, double-psi and swapped hairpin barrel, are classified together as the cradle-loop barrels due to the profile and shape of the binding pocket. Although the double-psi and swapped hairpin barrels are clearly related at the level of the βαβ-element, the relationship between the two folds is not obvious, as they cannot be interconverted by a simple topological transformation, such as circular permutation (VAT-Nnn and VAT-Nnc cp; Fig. 5.1). Thus, expression of VAT-Nnn(c) cp constructs resulted in not folded proteins. The addition of a beta strand to the ββαβ-element in the case of VAT-Nnc+βTa1217, which would resemble ββαββ-element of ABrB did give soluble but not folded protein, which leads us to believe that there are more sequence and structural determinants that connect these two folds. Existing construct could be used in directed evolution experiment in the search for the folded protein that would make strong link between these folds.

Bioinfomatical analysis in search for homologous of AbrB and subsequent clustering of the found sequences resulted in eight major groups of proteins (Fig. 4.17). The group of MraZ is the only group that contains two AbrB homologous sequences fused in one chain. MraZ is the first gene in the division and cell wall (dcw) cluster (Vincente et al, 1998) and our analysis suggests that it is a transcriptional regulator of the complete dcw cluster. Two halves of MraZ are nearest neighbors in the cluster map, suggesting that the protein originated through duplication and fusion of an ancestral homodimeric MraZ. The presence of the helical hairpin that connects two halves, also at the end of the second half, further sustains this hypothesis. The YjiW group is the only group that contains homodimeric and a single-chain version of AbrB. The single-chain version has as in the case of MraZ, two halves connected via a helical hairpin. Those proteins would be interesting from the structural point because the second copy lacks the first beta strand, which would suggest that the protein makes a seven stranded cradle-loop barrel.

Two other clusters of proteins Vir (with subcluster VagC) and PemI, are grouping proteins with similar functions in two different systems, plasmid maintenance and

toxin-antitoxin system (Katz et al., 1992; Zhang et al., 2003). Proteins from both groups, Vir and PemI, are transcriptional factors (antitoxins) but they come together in complex with different toxins. These toxins have either exonuclease or endonuclease activity, which is not enough elucidated. MazE protein (Fig 5.4) from the MazEF complex of the toxin-antitoxin system belongs to the group of PemI, and the presence of this protein in cluster map is of importance since it shows a clear homology with AbrB-N domain on the sequence and structural level.

Archeal PhoU sequences are found in phosphate-specific transport (pst) system operons, similar to the ones found in proteobacteria (Wanner, 1993). The N-terminal AbrB-like domain is connected to two PhoU-elements involved in phosphate uptake and it could regulate the pst operon directly since archaea lack the PhoRB two-component signal transduction system, which regulates the expression of the pst operon in bacteria. In a similar fashion two permeases of *Lactobacilli* could regulate cell homeostasis directly through the transcriptional regulation via the N-terminally positioned AbrB-like domain. The cyanobacterial protein cluster represents hypothetical sequences that are found to be chromosomally linked to cytochrome C subunits and riboflavin synthase.

5.1.3 SpoVT – transcriptional regulation through GAF domain

The core of the AbrB cluster is divided into the two recognizable subgroups: AbrB and SpoVT. The difference in the length of the sequence between AbrB and SpoVT, 90 compared with 180 residues, and different behavior of SpoVT and SpoVT-ΔC (Dong et al., 2004) prompted us to perform bioinformatical and structural analysis on this protein. It was shown that SpoVT protein is important in the transcriptional regulation of the last steps of spore formation in *B. subtilis* (Bagayan et al., 1996). This process is a series of cascades which are carefully regulated by multiple mechanisms. Prediction and later confirmation, by crystal structure of a GAF-domain at the C-terminal part of SpoVT is thus not so surprising, since it is known that GAF-domains tightly regulate function of other transcription factors (Babu and Teichmann, 2003). Regulation is exerted through a conformational change upon binding of the different small molecules to the GAF domain (Levdikov et al., 2006). Stringent

control of the last step in spore formation (Bagayan et al., 1996) is thus in good correlation with the structure of the SpoVT. The observed GTP binding affinity of SpoVT is also in the range of the binding affinity of the CodY protein (0.2-2mM), another transcription factor that regulates the transition from exponential growth to stationary phase in some gram positive bacteria (Ratnayake-Lecamwasam et al., 2001). Based on genome-coupling data a gene encoding a nucleoside triphosphate pyrophosphohydrolase (MazG) is present in every SpoVT operon. One could thus speculate that possible ligand for SpoVT GAF domain would be guanosine-tetraphosphate (ppGpp). A regulatory role of the ppGpp was already shown at different steps of the bacterial sporulation (Gentry et al., 1993), but this hypothesis for SpoVT has to be confirmed.
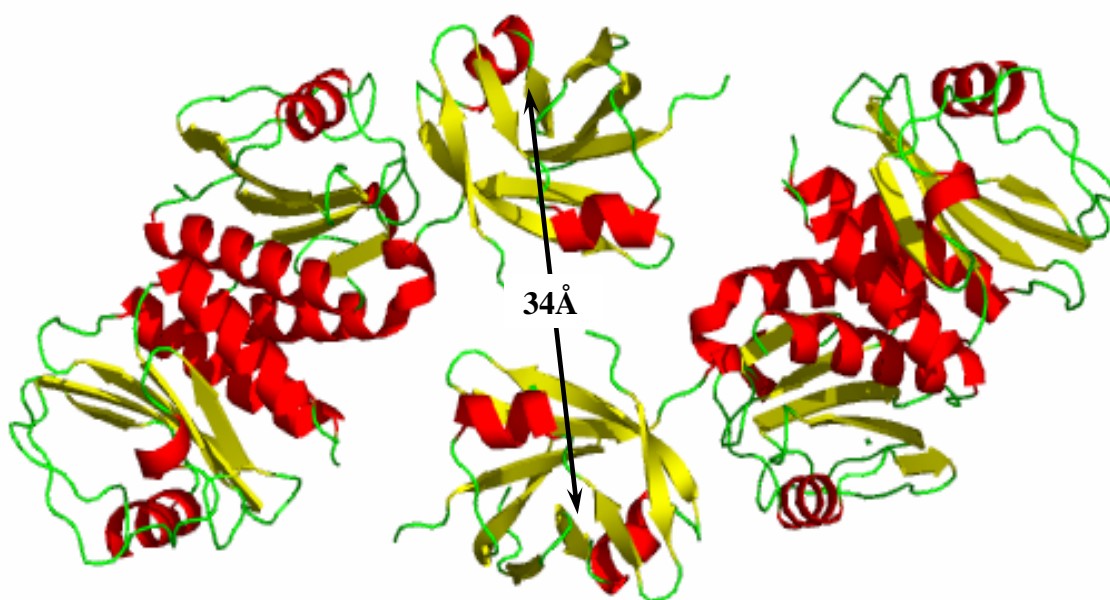
Figure 5.5 Structure of the SpoVT tetramer (top view into the DNA binding pockets). The arrow indicates two DNA binding domains and the calculated distance between them.

We have recently mapped a possible DNA binding sequence for SpoVT, which is a purin base rich palindrome-like sequence. The tetrameric structure of SpoVT gives us the possibility to propose that the role of the GAF domains in SpoVT is not only the regulation through 'small molecule' binding but also provides structural restrains for binding of such a DNA sequence. Since the dimerization of the N-terminal AbrB-like domain is forming one DNA binding site this would be sufficient for binding of only

one half of the palindrome and the presence of an additional dimer would be needed for complete binding of DNA. The tetrameric organization of SpoVT enables synchronized binding to a palindrome sequence and moreover dimerization of the GAF domains positions two DNA binding domains on ~34Å distance (Fig. 5.5). Importance of this distance is observed between two DNA binding domains in the structure of the catobolite activator protein (CAP) (Schultz et al., 1991) which binds similar palindromic sequences in two minor grooves of the DNA molecule. Ligand binding to SpoVT GAF domains might serve to fine-tune the orientation of the two DNA binding domains, as the binding of the DNA is not observed in absence of GTP. We are currently trying to co-crystallize SpoVT with the palindrome DNA sequence in presence of the GTP molecule.

5.1.4 RIFT barrels – origin of the cradle-loop barrels

The basis for the evolutionary model of the double-psi barrel (Vat-Nn), duplication of the βαβ-element and conservation of a GD-motif (GD box), was found to connect this fold with swapped hairpin barrels (AbrB) into a metafold named cradle-loop barrels (Fig. 5.4). An earlier proposed evolutionary path (Coles et al., 1999) connecting the two folds could not remain acceptable since inter-conversion of the folds by a simple topological modification, such as circular permutation was not possible. The NMR structure of the Phs018 protein from *P. horikoshii* (Fig. 5.6; Coles et al., submitted) showed a third topology of the duplicated conserved βαβ-element. The fold was named RIFT barrels, for its occurrence in **Ri**boflavin synthases, **F**1ATPase and **T**ranslation factors. The structure of Phs018 represents an intermediate between VAT-Nn and AbrB-N. It shares the six-stranded barrels topology with double-psi barrel, but without the psi-loop forming the crossover. In this respect it resembles features of the swapped hairpin barrels though it misses the additional beta strand per element that is found in this eight-stranded barrel (Fig. 5.6).
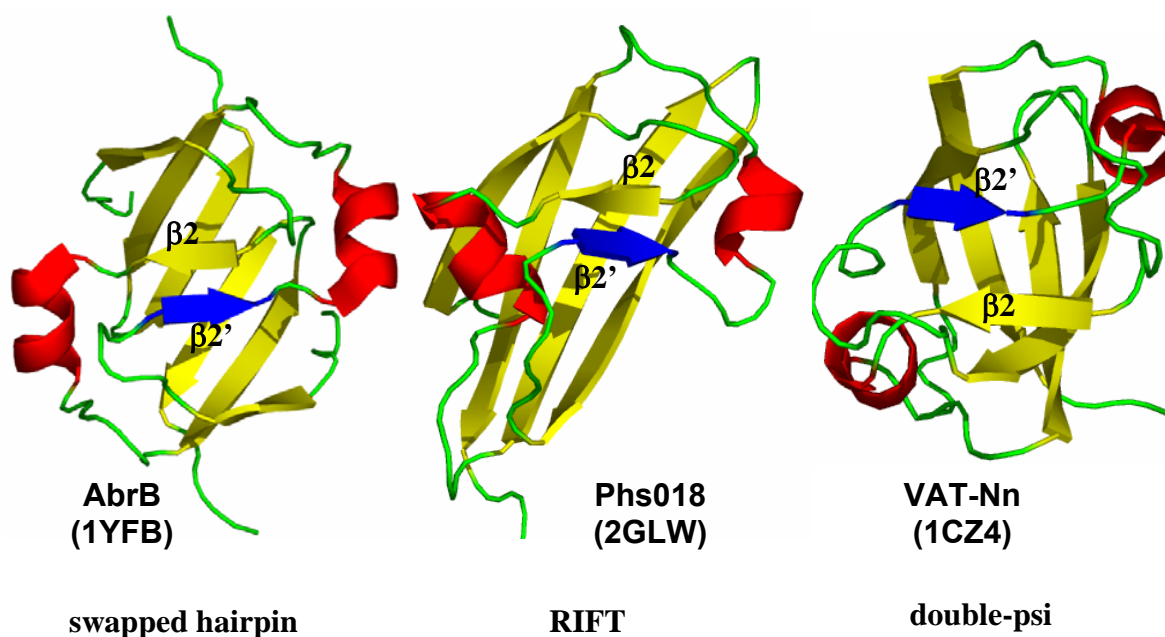
Figure 5.6 Comparison of the three different folds of cradle-loop barrels. Structures are orientated so that position of the β2' strand (blue) in respect to the β2 strand is easy to be seen. Notice the crossover of strands in VAT-Nn.

Even though we were unable to determine the complete structure of the Mt6002757 protein, limited NMR data show that a dimer of this protein assumes a fold of the Phs018 protein. Based on this data and conservation of the βαβ element in three folds it can be proposed that all of them arose from a sequence similar to the Mt6002757 protein, one canonical half of the RIFT barrel that forms a homodimer. The invasion of an additional β-strand at the C-terminus would result in the swapped hairpin fold of AbrB, which by further duplication and fusion could form a monomeric MraZ like protein (Fig. 5.2). Duplication and fusion of the Mt6002757-like sequence would result in a RIFT-like barrel, which is what we obtained by fusing two Mt6002757 sequences via a short two residue linker (Mt6002757 db). We were not able to determine the complete structure of this construct because of its instability, but limited NMR data shows that it assumes similar fold to the Phs018. The RIFT barrel fold would finally have given rise to the double-psi barrel through a strand swap of the first strand in each βαβ-element. CD spectroscopy data on the doubled loopless construct of VAT-Nnc (VAT-Nnc db loopless) and its overall stability make it

suitable for structure determination, which will give an answer to the importance of the length of psi-loops in the proposed strand swap event.

In many cases RIFT barrels have been substantially changed by insertions, usually followed by a loss of internal symmetry, which makes them hard to find on the sequence level. The Mj0056 protein is one of such examples, where insertion of the four elements in the first half of the barrel, made bioinformatics tools able to recognize only C-terminal canonical half of the protein. After structure determination of the Mj0056 it was obvious that the topology of the protein is similar to Phs018 with insertion of a helix in one of the cradle loops and three additional beta strands after the helix of the first half of the RIFT barrel (Fig. 5.8).
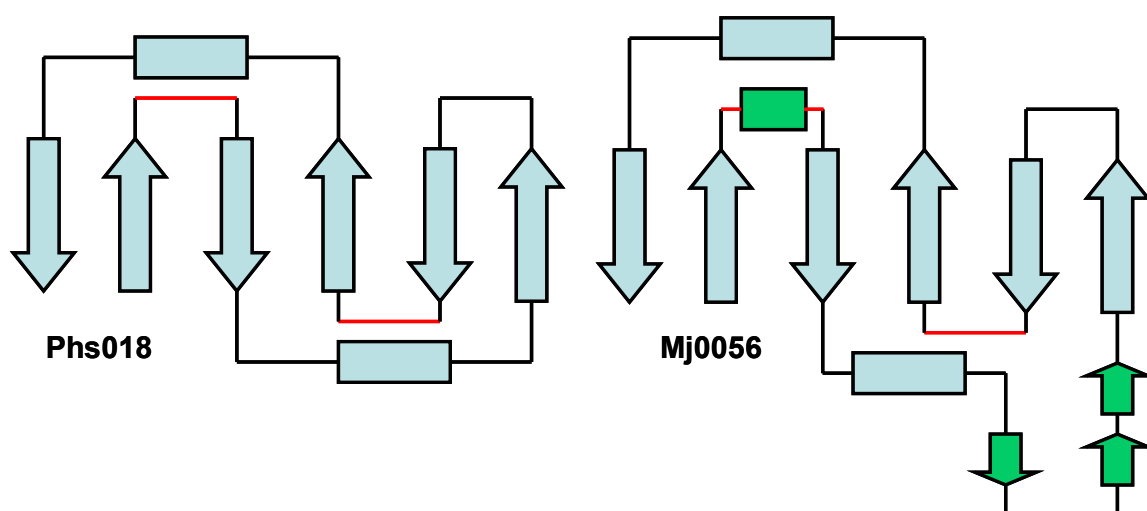
Figure 5.8 Topology of two RIFT barrels Phs018 and Mj0056. Elements inserted in the Mj0056 protein are colored in green; cradle-loops are colored in red.

Genome coupling data and annotation of the Mj0056 protein suggest that this protein is a transcriptional regulator of the riboflavin synthesis pathway, controlling the expression of the one of the rate limiting enzymes in this pathway, 3,4-dihydroxy-2-butanone-4-phosphate synthase (Richter et al., 1997). However we have not observed any unspecific DNA binding of this protein. Yet it might be that the protein needs to bind one of the precursors of riboflavin to actually bind to DNA. We are currently testing by NMR spectroscopy different compounds, from the riboflavin synthase pathway, for the possible binding to Mj0056. We are also trying to use the NMR structure for the molecular replacement in a new diffraction data set (2.7Å) collected from Mj0056 crystals in the search for possible bound ligand.

5.2 Beta-clam domains

Sequence similarity searches with AMA proteins yield the best matches to the ATPase domain of AAA metalloproteases. In a clustering study (Frickey and Lupas, 2004), it was shown that the AMA group is indeed closest to the metalloproteases and that phylogenetic reconstruction variously places it either at the root of this clade or even basally within it. Despite this, the ATPase domains of AMA sequences are clearly distinguished from those of metalloproteases by two main criteria, a gap of two residues in the 'second region of homology' (SRH), immediately adjacent to the arginine finger, and the presence of a sensor-2 arginine residue in the C-terminal helical subdomain. This residue makes contacts to the bound nucleotide and is present in most proteins of the AAA+ superfamily, but absent from virtually all AAA proteins. As was noted by Ogura et al. (2004), the presence of a sensor-2 arginine appears to exclude a second arginine in the SRH and indeed, AMA lacks the first arginine in the conserved SRH motif (RPGR) of AAA proteins (Appendix Fig. 4).

Most of the literature data show that AAA proteins oligomerize through their AAA modules (Vale, 2000, Ogura and Wilkinson, 2001) and in some cases they need ATP hydrolysis for the oligomerization process (Hartman and Vale, 1999, Babst et al., 1998). In contrast, AfAMA-ΔN, which comprises the AMA ATPase domain, did not form a higher oligomer and eluted as a dimer on gel-size exclusion columns. This unexpected finding suggests a pivotal role of the hexameric N-domain of AMA to the oligomerization and stability of the full-length ATPase complex. The recent structure of the cytoplasmic part of FtsH protein showing six-fold symmetry at C-terminal protease domain and three-fold symmetry at the level of AAA domain (Suno et al., 2006) would suggest that AMA would adopt a similar domain arrangement (Fig 5.9). This structure as well as our data on ATPase activity of the AfAMA and AfAMA-ΔN would definitely exclude the possibility of a concerted way of the ATP hydrolysis in those proteins, since maximally three nucleotide binding sites can be occupied at the same time. The structure of the FtsH AAA domain alone, forming spiral hexamers in presence of AMP-PNP (Niwa et al., 2002) or dimers in absence of nucleotides (Krzywda et al., 2002) would lead us to hypothesize about the role of the AMA N-terminal domain. This could be in indirect control of the ATPase activity of the AMA protein since the dimer of AfAMA-ΔN had a three fold higher consumption of the

ATP in comparison to the full length construct. AAA domains in AfAMA-ΔN construct would not be under stringent control of the N-domain hexamer, which would allow them to hydrolyze ATP at an elevated rate. This effect can not be seen in full AfAMA protein because of structural restraints opposed by the stable hexamer of the N-domain. More facts about the mechanism of ATP hydrolysis and the specificity of the AAA domain of AMA proteins will become evident from the structure of AfAMA.

p97 like proteins   **AMA and FtsH like proteins**
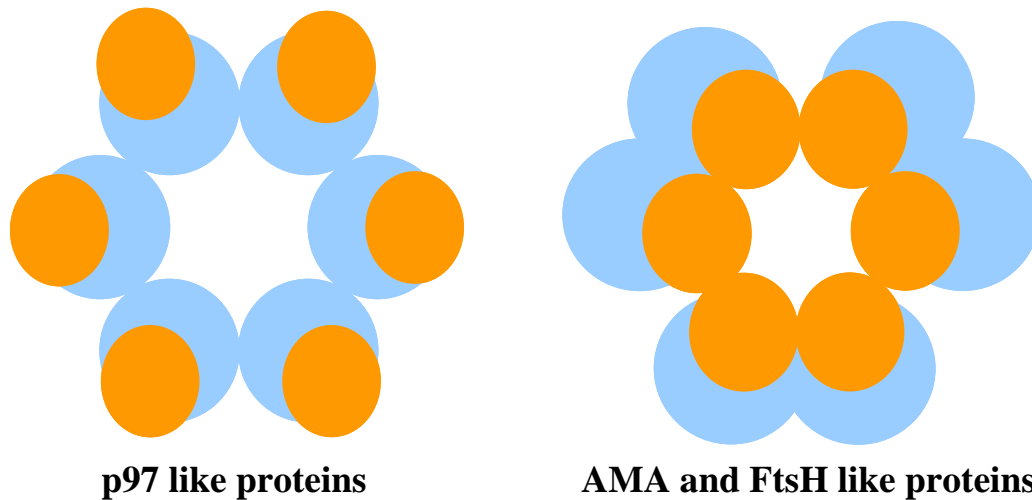
Figure 5.9 Two possible arrangements of the N-terminal (orange) and AAA domains (blue) in AAA proteins. The C-terminal protease domain would take the role of the N-domain in an FtsH model

The physiological role of AMA proteins is unknown. Yet, it is reasonable to assume that, like other AAA ATPases, AMA proteins are involved in protein unfolding and the dissociation of complexes, possibly in cooperation with other chaperones or proteases. We therefore tested AfAMA for its ability to interact with non-native proteins, as they may also be presented during unfolding of physiological substrate proteins. It was shown that both AfAMA and AfAMA-N can prevent aggregation of non-native substrates in heat induced aggregation assays, demonstrating a basic chaperone activity of AfAMA, which resides in the N-terminal domain. AfAMA-ΔN had no effect in these assays, indicating that the ATPase domain of AMA is not involved in substrate binding, at least for the proteins we tested. As the N-domain of AfAMA did not show ATPase activity, chaperone activity of the AfAMA protein is energy-independent. The coupling of ATP hydrolysis to substrate release from the N-domain is seen after addition of ATP to AfAMA-substrate complexes, which leads to

dissociation of the denatured substrate protein, prevention of rebinding and subsequent aggregation. This would imply that ATP hydrolysis is responsible for a conformational change in the N-domain, which induces the release of substrate. There are several known examples for intrinsic chaperone activity of AAA and AAA+ proteins (Arlt, et al. 1996, Golbik 1999, Benaroudj and Goldberg, 2000). However, there are conflicting data on the importance of N-domains in binding of protein substrates (Golbik et al., 1999; Lo et al., 2001; Beinker et al., 2002; Hinnerwisch et al., 2005). Chaperone activity of structurally related N-domains has been reported before for Cdc48-like and UFD proteins (Golbik et al., 1999; Park et al., 2005), where – in contrast to AMA – the β-clam subdomain occurs in context with an additional N-terminal double-psi barrel subdomain, which is supposed to share the ability to interact with non-native proteins. In the case of AMA proteins, the chaperone activity resides entirely within the β-clam like N-domain.

Comparison of the sequences of AMA N-domains with β-clam domains of p97 revealed the presence of a conserved sequence motif (GYPL), which resembles the pore motif in the ATPase domain of AAA+ proteins (Fig. 5.10; Yamada-Inagawa et al., 2003; Park et al., 2005). We did a detailed characterization of the functional and structural role of this conserved sequence motif and the loop, in which it is situated. We find that AfAMA requires an intact GYPL motif for hexamerization, as single or double mutations in this motif resulted in a loss of the oligomeric ring structure and formation of dimers. Moreover deletion of the motif ended in unfolded protein, but deletion of the eight residues following the GYPL motif in the same loop did not change the oligomeric state of the protein.

An intact GYPL motif is also important for the function of the AfAMA-N domain and the full AfAMA protein, since single or double mutants totally abolish chaperone activity. Similar mutations in the pore motif of FtsH (Yamada-Inagawa et al., 2003) or ClpX (Siddiqui et al., 2004) abolished proteolysis or substrate binding. It is important to note that the positions of these motifs are quite different in the context of the two AAA proteins.
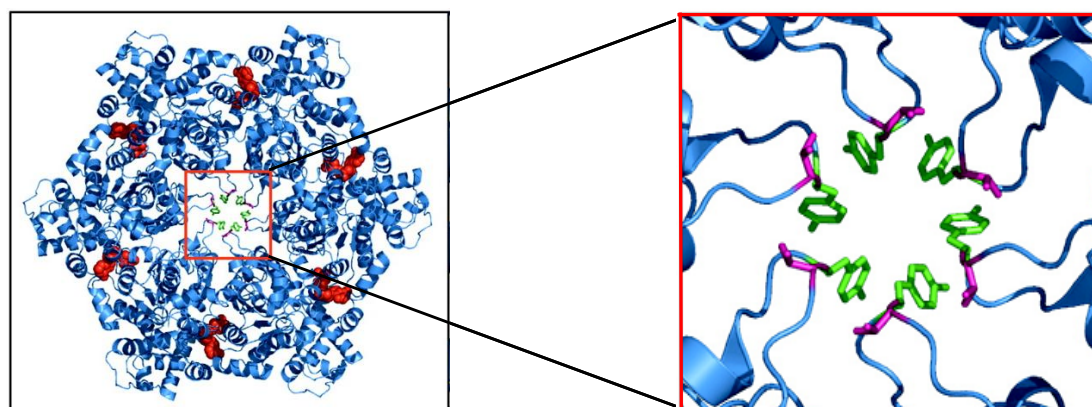
```
HslU    TKFTEVGYVGKEVKSI
ClpA    LIGAPPGYVGFDQGGL
ClpB    LVGAPPGYVGYEEGGY
ClpX    TTLTEAGYVGEDVENI
```

Figure 5.10 The pore motif in HslU, a member of the AAA+ superfamily. Alignment of GYVG motifs in ATPase domains of different AAA+ proteins

The GYPL motif is found in N-terminal β-clam domains of AMA proteins while GYVG is exclusively positioned in the AAA domains of several AAA+ proteins. It is noteworthy that insertion of the loop containing GYPL motif into the homologous sequence of the VAT-Nnc beta-clam induced oligomerization of this domain (monomer to trimer). Furthermore, trimeric Vat-Nnc chimera gained chaperone activity in heat aggregation assays with citrate synthase as a substrate. These data support both a structural and functional role of the GYPL motif.

Differences at the sequence level together with experimental data that we collected on the β-clam domain from Ph1500 protein (Ph1500N-monomer) and recently on a *Haloarcula maritsomi* protein (HmClam, gi: 55231035-trimer) indicate that the presence of a hydrophobic motif (AWI) and the length of the loop are important at least for oligomerization (Fig. 5.11), since none of these constructs had chaperone activity. Our hypothesis concerning the importance of the AWI motif, which is conserved in HmClam-like β-clams, needs to be further elucidated through mutational analysis and the structure(s) of AfAMA, MjAMA-N or Hm55231035.
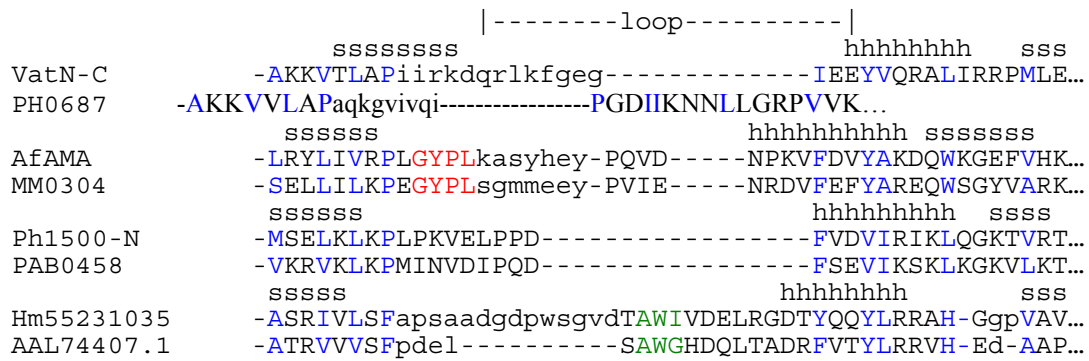
```
                              |--------loop----------|
                      sssssssss                       hhhhhhhh   sss
      VatN-C          -AKKVTLAPiirkdqrlkfgeg-------------IEEYVQRALIRRPMLE…
      PH0687        -AKKVVLAPaqkgvivqi----------------PGDIIKNNLLGRPVVK…
                      ssssss                        hhhhhhhhhh sssssss
      AfAMA           -LRYLIVRPLGYPLkasyhey-PQVD-----NPKVFDVYAKDQWKGEFVHK…
      MM0304          -SELLILKPEGYPLsgmmeey-PVIE-----NRDVFEFYAREQWSGYVARK…
                      ssssss                        hhhhhhhhh  ssss
      Ph1500-N        -MSELKLKPLPKVELPPD----------------FVDVIRIKLQGKTVRT…
      PAB0458         -VKRVKLKPMINVDIPQD----------------FSEVIKSKLKGKVLKT…
                      sssss                         hhhhhhhh     sss
      Hm55231035      -ASRIVLSFapsaadgdpwsgvdTAWIVDELRGDTYQQYLRRAH-GgpVAV…
      AAL74407.1      -ATRVVVSFpdel----------SAWGHDQLTADRFVTYLRRVH-Ed-AAP…
```

Figure 5.11 Alignment of the four different groups of β-clam domains. Proposed structural motifs, GYPL and AWI, are shown in red and green, respectively. Prediction of the secondary structure is indicated for each sequence; h-helix, s-strand. Conserved hydrophobic residues are colored blue. PH0687 (*P. horikoshii* gi: 14590568); MM0304 (*M. maze*i gi: 21226406); PAB0458 (*P. abyssi* gi: 14520879); AAL74407.1 (*H. volcanii* gi: 18539473)

Existing data on chaperone activity of β–clam domains in VAT-N (Golbik et al., 1999) and AMA (Djuranovic et al., 2006), or their interaction with other proteins like UFD1 (Park et al., 2005) and Ph1500N (this study) would indicate that these domains represent universal protein-protein interaction modules.

The difference in oligomerization and mode of the activity between those domains comes from the structural context in which they are found. Close connection with the double-psi barrel domain in VAT-N and UFD1 may have resulted in loss of the propensity for oligomerization and finally share of the function between both domains. This continues the discussion in the first part of this story where we showed the development of structure and chaperone activity of the double-psi barrel domain. One might speculate about a possible recruitment of a transcription factor with a RIFT-barrel like fold to an AMA-like beta-clam domain or a full AMA protein. Development of the RIFT barrel to the double-psi barrel fold was then followed by loss of AMA features in the beta-clam domain to result in UFD1 like proteins. Similar development but in context of a full AMA-like protein would result in an ancestral p97-like protein, which by duplication of the AAA module got today's structural

arrangement. This scenario is purely hypothetical, but is one of the possible scenarios in evolution of these AAA proteins based on our results.

Based on the proposed function and genome coupling data, the Ph1500 protein recruits or docks endonuclease III (Thayer et al., 1995) to sites of G-T mismatches induced by UV irradiation. According to the proposed mechanism, the hexamer of the C-terminal OB fold of Ph1500 slides down the DNA until it encounters a G-T mismatch, and the β-clam recruits endonucleas to fulfill its activity, which is excision of the mismatched pair. First results in which the beta-clam of Ph1500 co-elutes with endonuclease sustain this hypothesis, but further experiments have to be done to elucidate an interaction of Ph1500C domain with DNA and of Ph1500N with endonuclease.

5.3 PAN and ARC

The 26S proteasome, which is the major site of protein breakdown in mammalian cells, is composed of the 20S proteasome (molecular mass of 700 kDa) and two 19S regulatory complexes (700 kDa) (Coux et al., 1996; Baumeister et al., 1998). In the presence of ATP, the 19S complex associates with each end of the 20S cylinder and activate degradation of ubiquitinated and certain non-ubiquitinated proteins in an ATP-dependent process (Coux et al., 1996). The base of the 19S regulatory complex are six similar subunits which are members of the AAA family and were named RPT proteins (for Regulatory Particle Triple-A; Finley et al., 1998). Archaeal homologues of RPT subunits were also isolated and named PAN (for proteasome-activating nucleotidase; Zwickel et al., 1999). Several authors showed intrinsic chaperone activity of PAN and RPT in the absence of the proteasome (Braun et al., 1999; Strickland et al., 2000; Benaroudj and Goldberg, 2000) but no study mapped this activity to a certain domain of these proteins. Early studies on ARC from *R. erythropolis* predicted similarity between this protein and the ATPases of the eukaryotic 26S proteasome and their archaeal homologues of the PAN family (Wolf et al., 1998). These observations were usually based on the structural arrangement of the domains in these proteins, consisting of an N-terminal coiled-coil region followed by a β-strand rich interdomain and an AAA domain (Wolf et al., 1998; Zhang et al., 2004; Frickey and Lupas, 2004), and on genomics of *Actinobacteria* which lack PAN

sequences in their genomes. Since the available structural data on ARC and PAN proteins was limited to electron microscopy images (Wolf et al., 1998; Wilson et al., 2000; Zhang et al., 2004), we intended to further characterize those proteins by high-resolution methods. We were able to solve the crystal structure of the ARC-Nc domain, which revealed two almost identical OB folds arranged in a hexameric complex. The second OB fold could not be predicted by bioinformatics, since it contains an insertion of 20 residues in the loop connecting helix and β-strand 4' (Fig. 4.49). This insertion is not resolved in the structure as well. The central pore of the hexameric ARC-Nc ring has diameter of approx. 12 nm which resembles by size and shape the pore of most hexameric AAA proteins. Based on former biochemical and electron microscopy data on the ARC protein (Zhang et al., 2004), showing that the AAA domain of the protein is either dimeric or aggregated when expressed separately, it is likely that ARC-Nc is the hexamerization module of the ARC protein. Both OB folds seem to be needed for hexamerization, since single expression of any of the two OB folds in combination with the coiled-coil part or solely do not give hexameric complexes. The structure of ARC-Nc does not include the N-terminal coiled coil region which is found in both proteins, PAN and ARC. But we are eager to get it from newly obtained crystals of full ARC-N. We were not able to obtain any structure of the PAN-N protein or its parts, but homology of PAN-N and ARC-N indicates that PAN-Nc adopts the same structure like the first OB fold of ARC-Nc (Fig. 5.12). Even though similarity between these domains is striking, PAN-N does not form hexameric complexes. PAN-N domains tend to make high molecular complexes around 1MDa, which could be broken into dimers when expressed with an N-terminally fused thioredoxin. Apparently, the presence of the thioredoxin on the N-terminal coiled-coil region stabilizes dimers through sterical restrains that do not allow additional oligomerization of the coiled-coil region. The hexameric arrangement of the PAN-N subunits is restored in the context of the full PAN protein, thus showing that in this case the AAA module takes the role of the oligomerization interface.
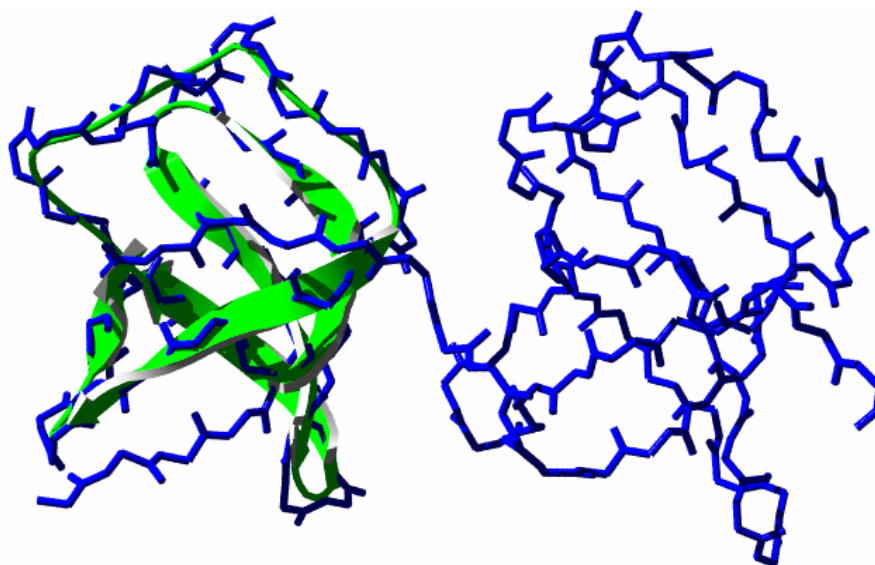
Figure 5.12 Ribbon model of the PAN-N OB fold (green) based on the crystal structure of ARC-Nc (blue).

Comparative analysis of the chaperone activity of the N-terminal domains of PAN and ARC showed that both domains are capable to prevent heat aggregation of different protein substrates: firefly luciferase and citrate synthase to high extent. The correct domain arrangement, N-terminal coiled-coil plus OB fold, was required for activity, since both subdomains did not show any activity when they were assayed by themselves. Earlier had been suggested that only the coiled-coil part is important for substrate recognition (Zhang et al., 2004; Reuter et al., 2004), but our results also indicate a prominent role for the OB fold domain. Taken together, our results further support the theory that N-terminal domains of AAA proteins are the primary substrate binding modules of AAA and AAA+ proteins (Golbik et al., 1999, Lo et al., 2001).

It is worth to note that ARC-N had in general a higher activity than the PAN-N construct. This might be explained by either the length of the coiled-coil part, 77 residues in ARC-N compared to 60 in PAN-N, or by the presence of the second OB fold in the structure of ARC–Nc, which increases stability of the ARC-N domain. Both scenarios are likely since shortening of the coiled-coils (PAN-N 3 to1 heptade constructs) or deletion of the second OB fold in ARC (ARC-N short) led to either reduced (PAN-N 3hep and PAN-N 2hep) or totally abolished activity (PAN-N 1 hep and ARC-N short) of these constructs.

Chimeras that were constructed from ARC-N and PAN-N parts (ARCPAN and PANARC) showed interchangeability of the coiled-coil and the OB fold subdomains.

Comparing the results among these chimeras (ARCPAN vs PANARC), and to the wild type N-terminal domains, it is noticeable that changes in the subdomains lead to the generation of proteins with differential affinity towards protein substrates. This can be attributed to the coiled-coil part of PAN-N and ARC-N domains, since this is the region with the most significant divergence. Several authors already implied that divergence of the coiled-coil region may be a determinant for the substrate specificity of PAN proteins (Zhang et al., 2004; Reuter et al., 2005; Medalia et al., 2006). Variability of the coiled-coil subdomain and its influence on the specificity and activity of these proteins was also confirmed by assaying "Frankenstein" proteins, where the coiled–coil of the wild type protein was substituted with the coiled-coil part of the GCN4 transcription factor from *S. cerevisae* (GARC, GPAN). Since the divergence of the OB folds in different PAN and ARC proteins is much lower then in coiled coil-part, we investigated also the importance of sequence conservation of the OB folds. The constructs AECS and ATIF, where the ARC-N coiled-coil is attached to OB folds from either cold shock protein of *E. coli* (AECS) or translation initiation factor from *M. janaschii* (ATIF), showed noticeable chaperone activity (up to 50% of ARC-N and PAN-N). Even more surprisingly, another 'frankenstein' construct, GECS, made completely from parts of proteins that are not related to ARC or PAN N-domains (coiled coil from GCN4 and OB fold from cold shock protein), still showed over 30% activity. The intriguing finding, that just by combining the RNA-binding OB fold (from either translation initiation factor or cold shock protein) with any coiled-coil can result in protein binding and chaperone activity, indicates that just a simple domain recruitment step was needed for these OB folds to change affinity towards a fundamentally different substrate, namely from RNA to protein.

Sequence analysis of the PAN and ARC N-terminal domains indicated that two proline residues in the hinge region between the coiled-coil and the OB fold(s) are highly conserved. The structure of the ARC-Nc revealed the position of only one of these prolines in each chain, which is buried between the OB folds of adjacent protomers. This happens in a way that rotation and isomerisation of this residue are very limited. Mutational analysis in which either one or both proline residues were substituted with alanine, greatly affected chaperone activity of both the wild type (ARC-N) and a "Frankenstein" protein (GPAN). Consequently, loss of both proline residues totally abolished chaperone activity of both constructs. In general, the activity was more reduced by mutation of the "buried" proline residue. Substitution of

that particular residue to the more bulky tryptophan resulted in a 3-fold higher chaperone activity of the "Frankenstein"-protein GPAN, but no obvious improvement or loss in the activity of wild type ARC-N. In summary, these results imply that strict orientation of the two subdomains, coiled-coil and OB fold, in ARC and PAN N-terminal domains has to be achieved for productive cooperation in substrate binding.

In the light of the presented results we are likely to propose that the surface area between the tips of the OB fold (first OB fold in ARC-N) and the coiled-coil subdomain is the substrate binding pocket in both ARC and PAN proteins. The position and sequence variability of the helix and loop connecting β3 and β4 strands in the predicted OB folds of different PAN and ARC proteins, which we noticed just recently, indicates that also this part is important for substrate binding. Yet, such hypothesis still waits to be experimentally confirmed.

6. Summary

AAA proteins are part of the large superfamily of AAA+ proteins, which are ringshaped P loop NTPases, whose common function is unfolding macromolecules in an energy-dependant manner. AAA proteins usually consist of an N-terminal domain, and one or two ATPase domains named D1 and D2. ATPase domains are relatively conserved within the family of AAA proteins and they are also thought to mediate hexamerization. N-terminal domains are important for substrate recognition and binding and, in contrast to the ATPase domains, they vary in their folds.

Based on published data and additional bioinformatic analysis of AAA proteins, we selected several different N-terminal domains from archaeal AAA proteins for functional and structural characterization. We also characterized proteins which share similar or related domains to the ones found in AAA proteins, making an important link between them. Heat and chemical aggregation assays of different substrate proteins were used to assay N-terminal domains, or full AAA proteins, for intrinsic chaperone activity. Protein structures were determined by crystallography or NMR spectroscopy.

Results of this study indicate that the barrel-like N-terminal domains of AAA proteins originated from the similar nucleic acid binding domains. A change in the affinity for substrate, from nucleic acid to protein, may have occurred through different mechanisms in the evolution. In the case of double-psi barrels this has probably happened through the evolution of a simple βαβ-motif found also in the RIFT and swapped hairpin barrels which are transcription factors, i.e. DNA binders. Structures of Mj0056 and SpoVT indicate that RIFT and swapped hairpin barrels have evolved further either by insertion of different structure elements (Mj0056) or by domain recruitment (SpoVT).

Similarity between the PAN and ARC N-domains was found to be both in structure and function. Both domains comprise a coiled-coil followed by one or two OB folds. OB stands for oligosaccharide binding and indicates that also this domain originated from a DNA-binding fold. Structure of the ARC-Nc subdomain and a comprehensive analysis of chimeric constructs of the coiled coils and OB folds indicate that ARC and PAN N-domains have arisen through evolution by domain

recruitment. Strict structural composition of the subdomains important in the chaperone function is maintained through the conserved PP-linker connecting the two subdomains.

Structural and functional characterization of the AfAMA, a member of the novel group of AMA AAA proteins, showed that substrate binding function and chaperone activity of these proteins resides in its β-clam like N-terminal domain. This domain can fulfill these functions independently, in contrast to the other homologous domains. We were able to show the importance of oligomerization for activity of these domains and that oligomerization is mediated by a small GYPL-motif found in a loop that presumably projects to the center of the hexamer. The N-terminal domain of AMA mediates hexamerization of the full protein independently from the AAA part of the protein and ATP utilization, which differs largely from other families of AAA proteins. Functional data on other β-clam domains: VAT-Nc, Ph1500N and Hm-clam would indicate that these domains represent universal protein-protein interaction modules
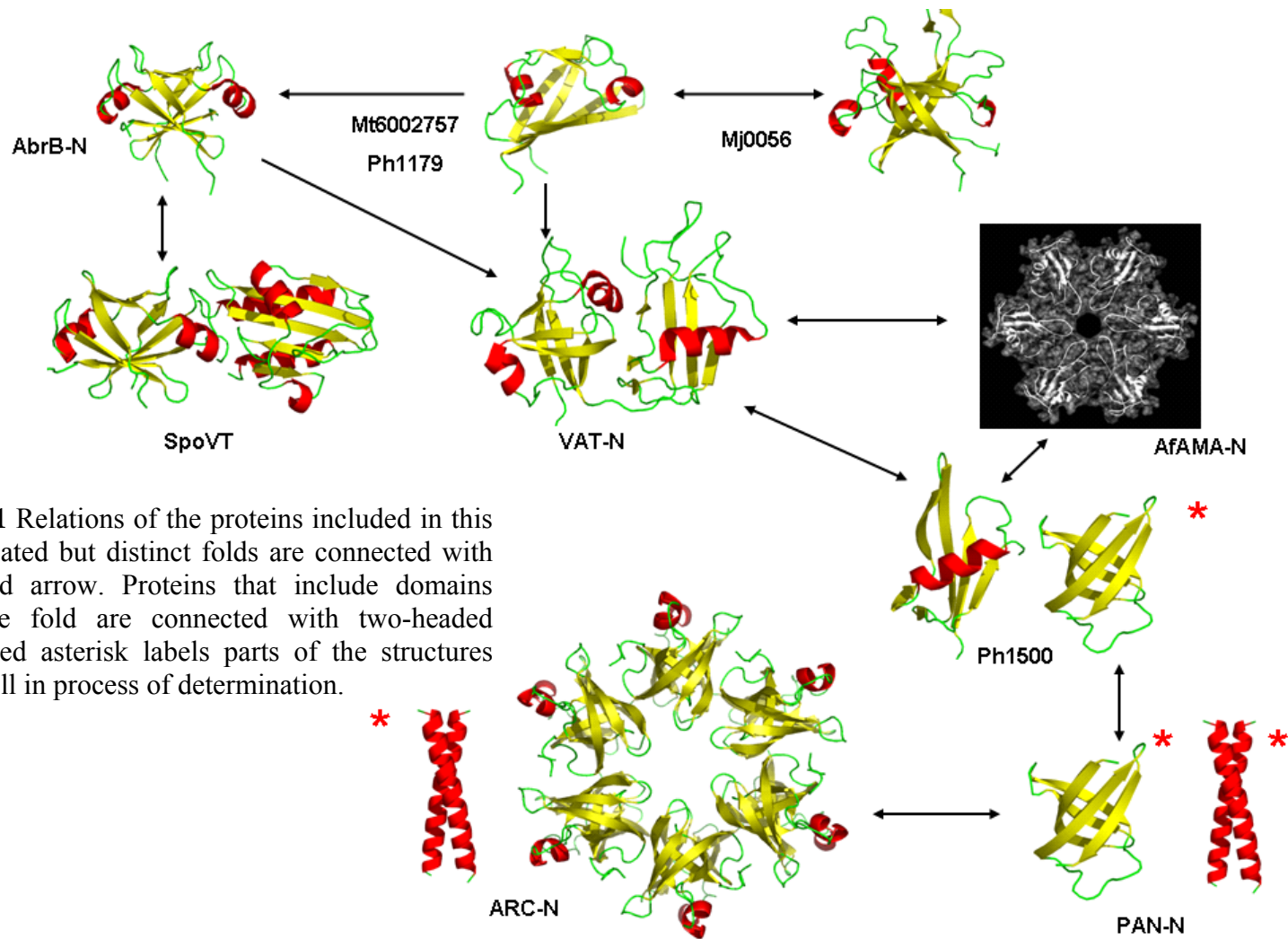
Figure 6.1 Relations of the proteins included in this study. Related but distinct folds are connected with one-headed arrow. Proteins that include domains with same fold are connected with two-headed arrows. Red asterisk labels parts of the structures that are still in process of determination.

7. Appendix
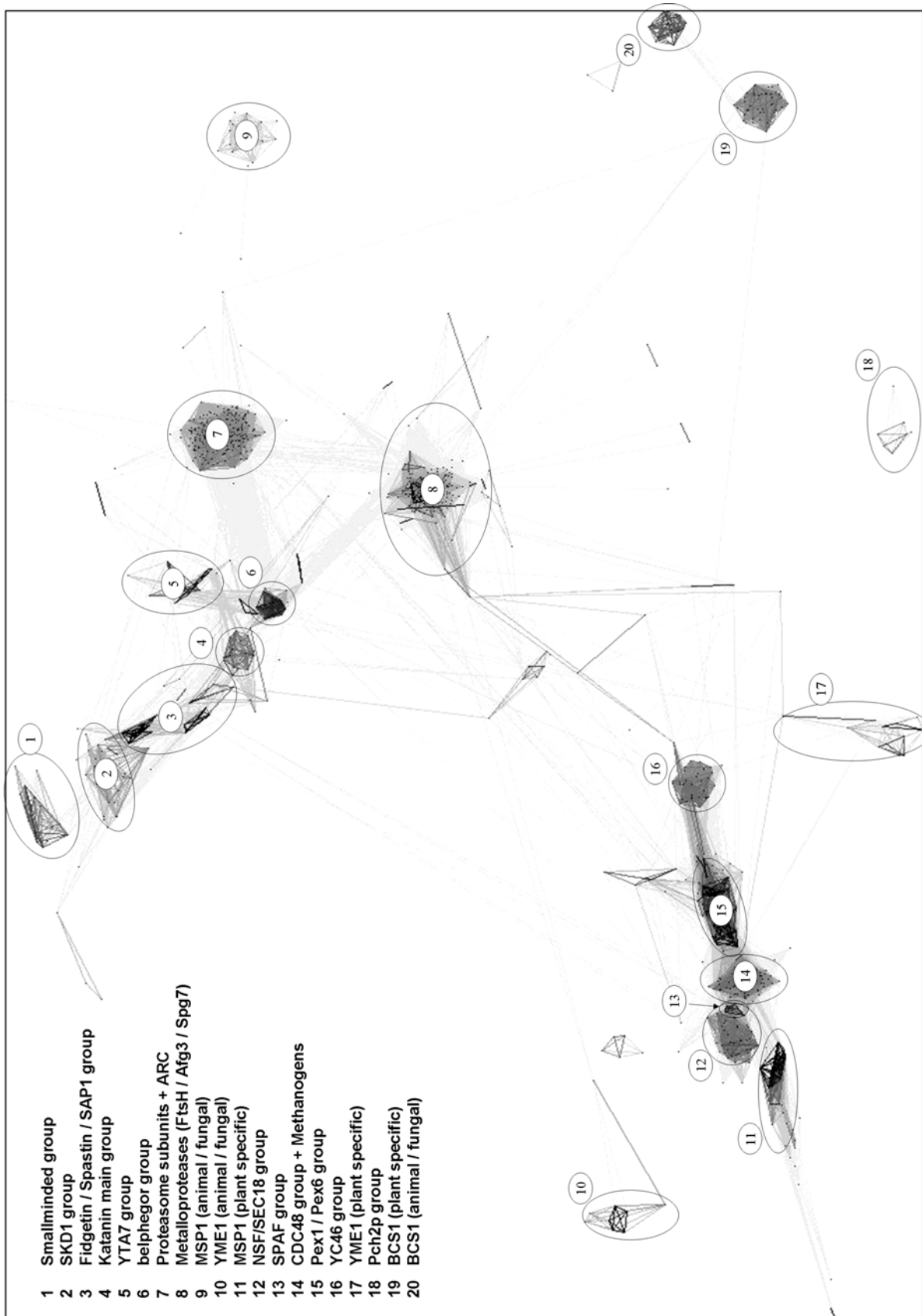
Figure 1. Clustering of the N-terminal domains of AAA proteins

1   Smallminded group
2   SKD1 group
3   Fidgetin / Spastin / SAP1 group
4   Katanin main group
5   YTA7 group
6   belphegor group
7   Proteasome subunits + ARC
8   Metalloproteases (FtsH / Afg3 / Spg7)
9   MSP1 (animal / fungal)
10   YME1 (animal / fungal)
11   MSP1 (plant specific)
12   NSF/SEC18 group
13   SPAF group
14   CDC48 group + Methanogens
15   Pex1 / Pex6 group
16   YC46 group
17   YME1 (plant specific)
18   Pch2p group
19   BCS1 (plant specific)
20   BCS1 (animal / fungal)

Figure 2. Evolution of the double-psi barrel showing the main postulated events along the branches

```
>gi|467445|B. subtilis         MKATGIVRRIDDLGRVVIPKEIRRTLRIREGDPLEIFVDRDGEVILKKYSPISELGDFAKEYADALYDSLGHSVL
>gi|52078551|B. licheniformis  MKATGIVRRIDDLGRVVIPKEIRRTLRIREGDPLEIFVDRDGEVILKKYSPISELGDFAKEYADALYDSMGHSVL
>gi|42779133|B. cereus         MKATGIVRRIDDLGRVVIPKEIRRTLRIREGDPLEIFVDRDGEVILKKYSPISELGDFAKEYADALYDSLGHNVL
>gi|30260245|B. anthracis      MKATGIVRRIDDLGRVVIPKEIRRTLRIREGDPLEIFVDRDGEVILKKYSPISELGDFAKEYAEALYDSLGHNVL
>gi|56378426|G. kaustophilus   MKATGIVRRIDDLGRVVIPKEIRRTLRIREGDPLEIFVDRDGEVILKKYSPISELGDFAKEYAEALFDSLGQPVL
>gi|22775742|O. iheyensis      MKATGIVRRIDDLGRVVVPKEIRRTLRIREGDPLEIFVDREGEVILKKYSPINELGHFAKEYAEALFQSLQTPVM
>gi|51858058|S. thermophilum   MKATGIVRRIDDLGRVVIPKEIRRTLRIREGDPLEIFVDRDGEVILKKYSPIGELGEFAKEYADSLYEAIGHTAL
>gi|68170176|D. hafniense      MKATGIVRRIDDLGRVVIPKEIRRTLRIREGDPLEIFVDREGEVILKKYSPIGELGDFAKEYADSLYEATGHIAC
>gi|77996477|C. hydrogenoformans MKATGIVRRIDDLGRVVIPKEIRRTLRIREGDPLEIFVDRNGEVILKKYSPIGELGDFAKEYAESLHESLGHIAF
>gi|68237484|M. thermoacetica  MKATGIVRRIDDLGRVVIPKEIRRTLRIREGDPLEIFVDREGEVILKKYSPIGELGDFAKEYADSLHEAIGHIAC
>gi|82746892|C. beijerincki    MKATGIVRRIDDLGRVVIPKEIRRTLRIREGDPLEIFTDRDGGVILKKYSPIEELTNFSKEYCESLQQVIGHVVL
>gi|82400554|C. saccharolyticus -----------------PKEIRRTLKIREGDPLEIYTDNEGEVILKKYSPIGEMGAFAKEYADTLHQVSGHIVI


>gi|467445|B. subtilis         ICDRDVYIAVSGSSKKDYLNKSISEMLERTMDQR-SSVLES-----DAKSV--QLVNGID—EDMNSYTVGPIVAN
>gi|52078551|B. licheniformis  ICDRDVYIAVSGGSKKDYLNKSISDMVEKAMDQR-TSVLDG-----DAKTI--QLIDGID—EEVASYTVGPIVAN
>gi|42779133|B. cereus         VCDRDSIIAVSGVSKKEYLNKSVGDLIEKTMEER-KSVIMT-----DESDI--SIIDGVT—EKVHSYTVGPIVAN
>gi|30260245|B. anthracis      VCDRDSIIAVSGVSKKEYLNKSVGDLIEKTMEER-KSVIMT-----DESDV--SIIDGVT—EKVHSYTVGPIVAN
>gi|56378426|G. kaustophilus   ICDRDVYIAVAGVSKKEYMNKSVSPLVEKAMEDR-NSILHT-----EEGEV--ELVDGMT—ETLKSYTIGPIVAN
>gi|22775742|O. iheyensis      ITDRDDVIAVAGESKKEYLNKPISNAIADTIEGR-SQVFEV-----DTKSM--EIIDGQE—QQLQSYCIHPVIAN
>gi|51858058|S. thermophilum   IADRDTIIAVAGAPKKEFLNKPIGSIVERAMEER-RSIVVSRAG--ETRQRGTIIGDDEDENRIPVYVVAPIVAG
>gi|68170176|D. hafniense      IADRDVIIAVSGASKKEYLNKPIWSSIEQAMAER-KTVQLK-----AGEGK--PDEDEEH—VKLTSQVAAPIIAE
>gi|77996477|C. hydrogenoformans IADRDTIIAVAGAPKKEFLNKPIGPAVERAMNER-KTVLITAT--GEHEYCKECITAEEGKCLFTSEIIAPIIAE
>gi|68237484|M. thermoacetica  IADRDNIIAVAGAPKKEFLDKPIGQAIERVMEER-KPLLANSP---SEDLF—PIDGEGEA-YKFTAEVIAPIIAE
>gi|82746892|C. beijerincki    IADKDAFVSVSGAPKKDYIERKVSSELEKIMDGR-KTVLLNKA---DNAVI-PLHNDDDE-TEYTSQVISPIIAE
>gi|82400554|C. saccharolyticus ITDRDKVIALSGASKKDYMDKALSQELERVMEENTIVFVKSEN---DMSSI--PIVEGDT—TRYTAQIVSPILSE


>gi|467445|B. subtilis         GDPIGAVVIFS--KDQTM—GEVEHKAVETAAGFLARQMEQ
>gi|52078551|B. licheniformis  GDPIGAVVIFS--KDQSI—GEVEHKAVETAAGFLARQMEQ
>gi|42779133|B. cereus         GDPIGAVIIFS--KEAII—SEIEHKAVNTAASFLAKQMEQ
>gi|30260245|B. anthracis      GDPIGAVIIFS--KEAII—SEIEHKAVNTAASFLAKQMEQ
>gi|56378426|G. kaustophilus   GDPIGAVIILS--REKTL-GEVEHKAVETAASFLARQMEQ
>gi|22775742|O. iheyensis      GDPIGCVLIFS--KEEKL-SKIEQKAAETASTFLAKQME-
>gi|51858058|S. thermophilum   GDPIGAVIICS—READAVMTETEIKLAETAAGFLAKQMEQ
>gi|68170176|D. hafniense      GDPIGAVILMS—KDPNVKMGDLELKLVETAAGFLAKQMEQ
>gi|77996477|C. hydrogenoformans GDPIGAVIIAS—REPGVKFGTLEQKMAETAAHFLAKQMEQ
>gi|68237484|M. thermoacetica  GDPIGAVILCS—REPGVKMGDMELKLAETAAGFLAKQMEQ
>gi|82746892|C. beijerincki    GDEIGAVIILS--KEDGELGEVETKLVETAAAFLGKQMEQ
>gi|82400554|C. saccharolyticus GTVIGSVIMCS—TESNVKMGDSEYKLVQAATSFFGKQLEQ
```
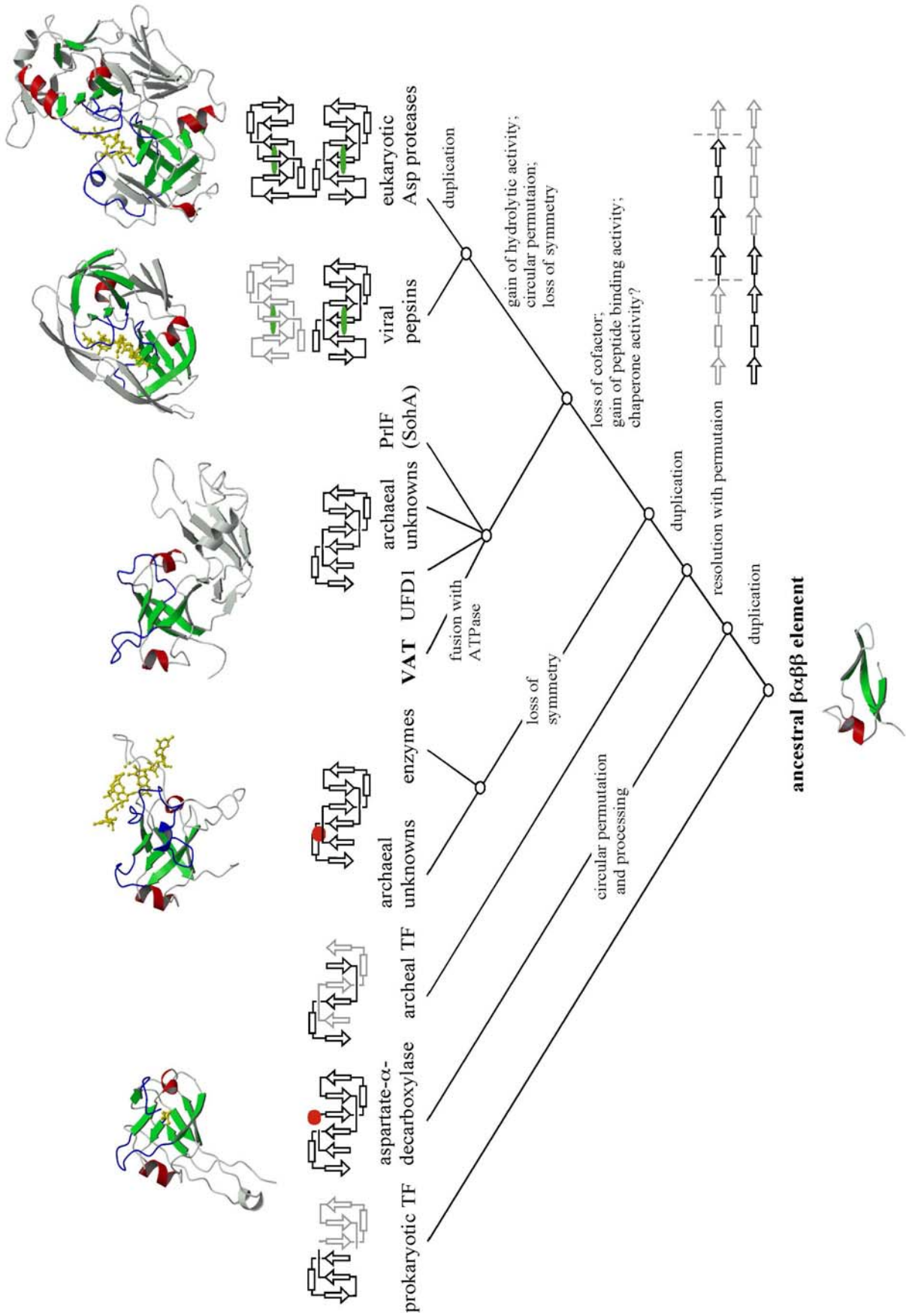
Figure 3. Alignment of SpoVT-like sequences from different sporulating bacteria. The AbrB-like and GAF domains are shown in red and black, respectively

```
                     |-------loop1---------|                        |---------
                    sss|ssssssss                 hhhhhhhh   sss  ssssss
VatN-C       -TEIAKKVTLAPIIRkdqrlkfgeg-------------IEEYVQRALIRRPMLEQDNISVpgltlagqtg
thvo10972    -TEIAKKVTLAPiirkdqrlkfgeg-------------IEEYVQRALIRRPMLEQDNISVpgltlagqtg
meth23963    -VKDAQKVVLAPvdqevi-------------------IRGDIRSAFLNRVLVKGDIIVSgirqhisggg
meja28672    -IKEAKKVVLAPtqpirfgpg----------------FEDFVKRKILGQVLSKGSKVTIgvlgta----
aful3736     -AKPAEKVTLAPtepvrlmg-----------------GEAYLLRLLEGRPVIKGQKIRVevfght----
aful4530     -YQPAKTVILAPlkkmdlriygvd-------------IGEYLKHQFLKRPVVEGDLVPLvgspalsgfg
pyho363348   -VKEARKVVLAPtepirfgrd----------------FVEWLHERLVGRPVVRGDYIKIgvlgqe----
pyho362195   -VKEAKKVVLAPaqkgvivqi----------------PGDIIKNNLLGRPVVKGDIVVAsgrgdlyyss
hasp30078    -AEKADTLVLAPpeeasvqfgsd--------------AAGMVKRQILKRPVVARDIVPVmsstnhpfmr
hasp31046    -VEPADRVTVSLpqnlqirgd----------------LGSHLREHLVDQAVRAGQTVAFpigfgmfsgr
hasp29931    -IKPAGGVTVALpqnlrvrgn----------------IAPMVRDRLNGRPVTAGQTIPIsfgfggmsti
suso10166    -VKPASTVKLAPsnfsitvdpg---------------FISYVKKRLKEFPLVEGDTVLIpvlgqa----
suso10402    -VQDATKVVLAPtqpisfsqs----------------FVEYVKDWLMDKPISRGETISVptyvgs----
aepe423790   kVEPASKVVLAPtepirfgrd----------------FVEYVKEFLLRKPISRGETIIVpvleg-----
aepe423106   -VRTATKVKLAPvsytmtvdeg---------------FKRYVKKKLQGVPITEGDVVVVpvigqa----
pyba21772    -PKPAAFVKLAPvsmtiavdan---------------FLQYIKQRLREYVLVEGDMLQIhvlsqp----
pyba23306    -LKPAQRVVLTPtepvrvd------------------SEYLKKQILLGKPVARGQAIDVpfygga----
                  ssssss     ssss                hhhhhhhhh ssssss    ssss
AfAMA        ----LRYLIVRPLGYPLkasyhey-PQVD-----NPKVFDVYAKDQWKGEFVHKNKLIFDmrmfpd----
PRS2_METJA   ----LKYVVLEPAGFPIrvssenvkVSTD-----DPILFNIYARDQWIGEIVKEGDYLFDnsilpd----
PRS2_METTH   ----AKLVVLQPVGYPFvcnlmea-PRIDav---NKELFEIYARDQWEGFRAAEGSYLFDqkllpd----
MM0304       ----SELLILKPEGYPLsgmmeey-PVIE-----NRDVFEFYAREQWSGYVARKGDYLFDrrmfpd----
MK1368       -----------MGYPVrepgmgkeVVVD-----SLEAFNAYAREQWLGEVVREGTILFDtgvvhs----
                  ssssss           hhhhhhhhh ssss   sssssss
Ph1500-N     ----MSELKLKPLPKVELPPD----------------FVDVIRIKLQGKTVRTGDVIGISILGKE----
H75108       ----VKRVKLKPMINVDIPQD----------------FSEVIKSKLKGKVLKTGDVVSVDILGKE----
AAL81354.1   ----NMMIILKPLIEVDLPED----------------FVDIIKAKLKGQTVKSGEIITVDILGKP----
                  sssss               ssss  hhhhhhhh     hhhhhhh
AAL74407.1   ----ATRVVVSFpdel----------SAWGHDQLTADRFVTYLRRVHEd--AAPGDEWEEFLDVGCcgds
AAV47916.1   ----ASRVELSYpadl----------GWGRDKIDGSPFRAYLRKTHNt--ATPGDVWTEFVGVGCcgdt
AAV46454.1   ----ADRIVLSFapsaadgdpwsgvdTAWIVDELRGDTYQQYLRRAHGgp-VAVGEEWAEFVSCGCatpq


                     ---loop2-----------|
                               ssssssss    sss   ssssss
VatN-C       -------------------LLFKVVKTLPSKVpVEIGEETKIEIREEPASevleevsr----------
thvo10972    -------------------LLFKVVKTMPGKVpVEIGEETKIEIREEPASevleevsr----------
meth23963    lfdeffrdfmdisplge----IKLAVVSTSPAGV-VRVTPSTQVEMQQKPVDvsklegvknlvd------
meja28672    -------------------LTFVVVSTTPAGP-VRVTDFTHVELKEEPVSeiketkvpd---------
aful3736     -------------------LTFVITATRPSGV-VVVTRNTAIELKEKPAEeevkravpd----------
aful4530     rynqqnqa------------VVFVAVKTEPKGP-VVIDETTKVVYRDRPAKgferfgkag---------
pyho363348   -------------------LTFVVTTTQPSGV-VQITEYTDFDISEKPVKevekrmttg---------
pyho362195   gtpfdeifrgffeamsvgfgeLKFMVVNTIPKGI-VQITYNTEVEVLPQAVEvreekipe----------
hasp30078    spgqa---------------IPLIAVETEPEGV-CLVTEDTDVELREEPISgfertggg----------
hasp31046    sgrr----------------IPLRVVDTQPSGT-VVVQNTTEIEIADQSAQevsvesgepenttapa--
hasp29931    sgqq----------------IPVKIAETEPSGT-VVVSNDTEIQLSERPAEeiapgageaaetgdptpn
suso10166    -------------------IPFTVVQVKPAGI-VLVNDDTIISISDKPVEpsrypr------------
suso10402    -------------------IDFVVVSTQPSQS-VRITGRTSLEIRQEPVKesaavpk-----------
aepe423790   -------------------LPLVVVSTQPAHF-VYVTEATEVEIREKPVReeierlrgvpk-------
aepe423106   -------------------VQLQVVDARPKGA-VIVSEETIVDVLEKPVAqsrvpk------------
pyba21772    -------------------LTFQVVQTKPSNTvLIITEDTQIQIFEKPVSgvkiph------------
pyba23306    -------------------IRFVVVQVQPGPA-AYVSIDTEVTVREEPVKeaeltipr----------
                               ssssssss    ss  sssssss
AfAMA        -------------------FAFEVIDCDPps--GYISDSTIILVESDPRVieteivrd----------
PRS2_METJA   -------------------YAFKVISTYPkeg-GMITSETVFKLQTPKKVlrtqfkk-----------
PRS2_METTH   -------------------YAFKIIRAHPdg--SKITRNTSIILLENDREefhevrss----------
MM0304       -------------------FAYRIIDVEPae--SMIGSSTSIIVTEEESGipssseikgd--------
MK1368       -------------------YAFKVVRVVPsgm-GRITSSTRFVLRTRFEEdrmeipn----------
                               sssssss    sss   ssssss
Ph1500-N     -------------------VKFKVVQAYPSP--LRVEDRTKITLVTHP
H75108       -------------------IRFKVVQAMPSP--LTVDESTGVLLTRHS
AAL81354.1   -------------------IEFKVLYAEPSP--VKVTQKTQIKFAKGN
                               ssssssss    ss  sssssss           ssss
AAL74407.1   l-----------------TLTLRVEELDPEGT-TRIDDETTVEFVEREgs-VHGGWCVQSADGPvs
AAV47916.1   l-----------------DFPLQVESV--EGG-SAVTEDTEFVYTEREacgIAGGWQVQSAAGPte---
AAV46454.1   ------------------DVVLRVERI--EGG-MALGDATTLAIHPrndaetap--------------
             -
```

Figure 3. Alignment of different β-clam domains. p97/CDC48 (green), AMA (blue), Ph1500-like (red) and Hm-like (black). The secondary structure prediction is written above each group.

7.1 Sequences of the proteins

Thermoplasma acidophilum VAT ATPase (VCP-like ATPase)
GI:16081896

MESNNGIILRVAEANSTDPGMSRVRLDESSRRLLDAEIGDVVEIEKVRKTVGRVYRARPEDEN
KGIVRIDSVMRNNCGASIGDKVKVRKVRTEIAKKVTLAPIIRKDQRLKFGEGIEEYVQRALIR
RPMLEQDNISVPGLTLAGQTGLLFKVVKTLPSKVPVEIGEETKIEIREEPASEVLEEVSRISY
EDIGGLSEQLGKIREMIELPLKHPELFERLGITPPKGVILYGPPGTGKTLIARAVANESGANF
LSINGPEIMSKYYGQSEQKLREIFSKAEETAPSIIFIDEIDSIAPKREEVQGEVERRVVAQLL
TLMDGMKERGHVIVIGATNRIDAIDPALRRPGRFDREIEIGVPDRNGRKEILMIHTRNMPLGM
SEEEKNKFLEEMADYTYGFVGADLAALVRESAMNALRRYLPEIDLDKPIPTEILEKMVVTEDD
FKNALKSIEPSSLREVMVEVPNVHWDDIGGLEDVKREIKETVELPLLKPDVFKRLGIRPSKGF
LLYGPPGVGKTLLAKAVATESNANFISIKGPEVLSKWVGESEKAIREIFKKAKQVAPAIVFLD
EIDSIAPRRGTTSDSGVTERIVNQLLTSLDGIEVMNGVVVIGATNRPDIMDPALLRAGRFDKL
IYIPPPDKEARLSILKVHTKNMPLAPDVDLNDIAQRTEGYVGADLENLCREAGMNAYRENPDA
TSVSQKNFLDALKTIRPSVDEEVIKFYRTLSETMSKSVSERRKQLQDQGLYL

*VAT-Nn underlined

Methanothermobacter thermautotrophicus excisionase-like protein
GI:6002757

MAMEDVGVPFSNRLTRQGNIKVPADLRDALKLKPGDLLVVEIKKVDRS

Pyrococcus horikoshii Ph1179 GI:3257596

MQNQQKTVEPLAKFHASVNIKGQLVVPVKDREVFGLKRGDILEIIVRSFDVINGKIHIKKRAY
ILVRLSSKGLITIPEEVRRELGISPGDTVEVLLVGFHKFDELVTEKGKQIAKLIQANTHMRLI
TSEEEKTIIEKSRTYYV

Thermoplasma acidophilum Ta1217 GI:16082225

MTDNKKIMDIARMTKRGASVRVTIPKKVLKKLNFKDEDLIAFYESEDGRIYIDLLK

Methanocaldococcus jannaschii Mj0056 GI:2128102

MVKLMIIEGEVVSGLGEGRYFLSLPPYKEIFKKILGFEPYEGTLNLKLDREFDINKFKYIETE
DFEFNGKRFFGVKVLPIKILIGNKKIDGAIVVPKKTYHSSEIIEIIAPMKLREQFNLKDGDVI
KILIKGDKDE

Pyrococcus horikoshii Ph1500 GI:3257925

MEGVIMSELKLKPLPKVELPPDFVDVIRIKLQGKTVRTGDVIGISILGKEVKFKVVQA
YPSPLRVEDRTKITLVTHPVDVLEAKIKGIKDVILDENLIVVITEENEVLIFNQNLEELY
RGKFENLNKVLVRNDLVVIIDEQKLTLIRT

Bacillus subtilis AbrB GI: 113009

MFMKSTGIVRKVDELGRVVIPIELRRTLGIAEKDALEIYVDDEKIILKKYKPNMTCQV
TGEVSDDNLKLAGGKLVLSKEGAEQIISEIQNQLQNLK

```
*Abrb-N underlined
```
Methanococcus jannaschii AMA Mj1494 GI: 3915816

MSKIGFNPIKIKSFSKIKTYDDTLPSLKYVVLEPAGFPIRVSSENVKVSTDDPILFNIYARDQ
WIGEIVKEGDYLFDNSILPDYAFKVISTYPKEGGMITSETVFKLQTPKKVLRTQFKKAKFSEI
IGQEEAKKKCRIIMKYLENPKLFGEWAPKNVLFYGPPGTGKTLMARALATETNSSFILVKAPE
LIGEHVGDASKMIRELYQRASESAPCIVFIDELDAIGLSREYQSLRGDVSEVVNALLTELDGI
KENEGVVTIAATNNPAMLDPAIRSRFEEEIEFKLPNDEERLKIMELYAKKMPLPVKANLKEFV
EKTKGFSGRDIKEKFLKPALHRAILEDRDYVSKEDLEWALKKILGNRREAPQHLYL

```
*MjAMA-N is underlined
```

Archaeoglobus fulgidus AMA AF1285 GI: 2649294

MAKRETAELRYLIVRPLGYPLKASYHEYPQVDNPKVFDVYAKDQWKGEFVHKNKLIFDMRMFP
DFAFEVIDCDPPSGYISDSTIILVESDPRVIETEIVRDITLDDVVGQEEAKRKAKVILEYLRN
PEKFGKWAPKNVLFYGPPGTGKTMMAKALSNEAKTPFLSVKSTKLIGEHVGDGARRVHELYER
ARQLAPCIVFLDEFDAIALDRGYQEIRGDVSEIVNALLTELDGTNSNEGICTIAATNRVELLD
ASIRSRFEEEIEFRLPSYEERLEILRRNLEEFPVPVKARLELVAAASEGFSGRDLVEKVIKAS
LHKAIAEGKDKIETEDLLTAAEKVKMPSRQPPKQMFV

*AfAMA-N is underlined

Archaeoglobus fulgidus PAN Af1976 GI:3122632

MGDSEIQYLLEKLKKLEEDYYKLRELYRRLEDEKKFIESERIRYEREVRRLRSEVERL
RSPPLLVGVVSDILEDGRVVVKSSTGPKFVVNTSQYINEEELKPGARVALNQQTLAIV
NVLPTSKDPMVYGFEVEEKPEVSYEDIGGLDVQIEEIREAVELPLLKPELFAEVGIEPP
KGVLLYGPPGTGKTLLAKAVANQTRATFIRVVGSEFVQKYIGEGARLVREVFQLAKE
KAPSIIFIDELDAIAARRTNSDTSGDREVQRTMMQLLAELDGFDPRGDVKVIGATNRI
DILDPAILRPGRFDRIIEVPLPTFEGRIQIFKIHTRKMKLAEDVDFKELARITEGASGADI
KAICTEAGMFAIREERAKVTMLDFTKAIEKVLKKTTPIPDLKGVMFV

*AfPAN-N is underlined

Methanocaldococcus jannaschii PAN Mj1176 GI:2492524

MVFEEFISTELKKKEKKAFTEEFKEEKEINDNSNLKNDLLKEELQEKARIAELESRILKL
ELEKKELERENLQLMKENEILRRELDRMRVPPLIVGTVVDKVGERKVVVKSSTGPSF
LVNVSHFVNPDDLAPGKRVCLNQQTLTVVDVLPENKDYRAKAMEVDERPNVRYEDI
GGLEKQMQEIREVVELPLKHPELFEKVGIEPPKGILLYGPPGTGKTLLAKAVATETNA
TFIRVVGSELVKKFIGEGASLVKDIFKLAKEKAPSIIFIDEIDAIAAKRTDALTGGDREV
QRTLMQLLAEMDGFDARGDVKIIGATNRPDILDPAILRPGRFDRIIEVPAPDEKGRLEI
LKIHTRKMNLAEDVNLEEIAKMTEGCVGAELKAICTEAGMNAIRELRDYVTMDDFR
KAVEKIMEKKKVKVKEPAHLDVLYR

*MjPAN-N is underlined
Rhodococcus erythropolis ARC GI:3790601

MSSTENPDSVAAAEELHALRVEAQVLRRQLAQSPEQVRELESKVDSLSIRNSKLMDT
LKEARQQLIALREEVDRLGQPPSGYGVLLSVHEDKTVDVFTSGRKMRLTCSPNIDTD
TLALGQTVRLNEALTIVEAGTYEQVGEISTLREVLDDGLRALVVGHADEERIVWLAA
PLAAVFADPEADIIAYDADSPTRKLRPGDSLLVDTKAGYAFERIPKAEVEDLVLEEVP
DVHYDDIGGLGRQIEQIRDAVELPFLHKDLFHEYSLRPPKGVLLYGPPGCGKTLIAKA
VANSLAKKIAEARGQDSKDAKSYFLNIKGPELLNKFVGETERHIRMIFQRAREKASEG
TPVIVFFDEMDSIFRTRGSGVSSDVETTVVPQLLSEIDGVEGLENVIVIGASNREDMIDP
AILRPGRLDVKIKIERPDAESAQDIFSKYLVDGLPINADDLAEFGGDRTACLKAMIVR
VVDRMYAESEENRFLEVTYANGDKEVLFFKDFNSGAMIQNIVDRAKKYAIKSVLDT
GAPGLRVQHLFDSIVDEFAENEDLPNTTNPDDWARISGKKGERIVYIRTLVTGKNASA
SRAIDTESNTGQYL

*ARC-N is underlined

Haloarcula marismortui Hm clam GI:55231035

MAPVADRIVLSFAPSAADGDPWSGVDTAWIVDELRGDTYQQYLRRAHGGPVAVGE
EWAEFVSCGCATPQDVVLRVERIEGGMALGDATTLAIHPRNDAETAP

8. References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res* **25**(17): 3389-402.

Arlt, H., Tauer, R., Feldmann, H., Neupert, W. and Langer, T. (1996). "The YTA10-12 complex, an AAA protease with chaperone-like activity in the inner membrane of mitochondria." *Cell* **85**(6): 875-85.

Babst, M., Wendland, B., Estepa, E. J. and Emr, S. D. (1998). "The Vps4p AAA ATPase regulates membrane association of a Vps protein complex required for normal endosome function." *Embo J* **17**(11): 2982-93.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. and Teichmann, S. A. (2004). "Structure and evolution of transcriptional regulatory networks." *Curr Opin Struct Biol* **14**(3): 283-91.

Bagyan, I., Hobot, J. and Cutting, S. (1996). "A compartmentalized regulator of developmental gene expression in Bacillus subtilis." *J Bacteriol* **178**(15): 4500-7.

Baumeister, W., Walz, J., Zuhl, F. and Seemuller, E. (1998). "The proteasome: paradigm of a self-compartmentalizing protease." *Cell* **92**(3): 367-80.

Beinker, P., Schlee, S., Groemping, Y., Seidel, R. and Reinstein, J. (2002). "The N terminus of ClpB from Thermus thermophilus is not essential for the chaperone activity." *J Biol Chem* **277**(49): 47160-6.

Benaroudj, N. and Goldberg, A. L. (2000). "PAN, the proteasome-activating nucleotidase from archaebacteria, is a protein-unfolding molecular chaperone." *Nat Cell Biol* **2**(11): 833-9.

Beyer, A. (1997). "Sequence analysis of the AAA protein family." *Protein Sci* **6**(10): 2043-58.

Braun, B. C., Glickman, M., Kraft, R., Dahlmann, B., Kloetzel, P. M., Finley, D. and Schmidt, M. (1999). "The base of the proteasome regulatory particle exhibits chaperone-like activity." *Nat Cell Biol* **1**(4): 221-6.

Castillo, R. M., Mizuguchi, K., Dhanaraj, V., Albert, A., Blundell, T. L. and Murzin, A. G. (1999). "A six-stranded double-psi beta barrel is shared by several protein superfamilies." *Structure* **7**(2): 227-36.

Chen, S., Jancrick, J., Yokota, H., Kim, R. and Kim, S. H. (2004). "Crystal structure of a protein associated with cell division from Mycoplasma pneumoniae (GI: 13508053): a novel fold with a conserved sequence motif." *Proteins* **55**(4): 785-91.

Coles, M., Diercks, T., Liermann, J., Groger, A., Rockel, B., Baumeister, W., Koretke, K. K., Lupas, A., Peters, J. and Kessler, H. (1999). "The solution structure of VAT-N reveals a 'missing link' in the evolution of complex enzymes from a simple βαββ element." *Curr Biol* **9**(20): 1158-68.

Coles, M., Djuranovic, S., Soding, J., Frickey, T., Koretke, K., Truffault, V., Martin, J. and Lupas, A. N. (2005). "AbrB-like transcription factors assume a swapped hairpin fold that is evolutionarily related to double-psi beta barrels." *Structure* **13**(6): 919-28.

Coles, M., Hulko, M., Djuranovic, S., Truffault, V., Koretke, K., Martin, J. and Lupas, N. A. (2006). "Common evolutionary origin of swapped-hairpin and double-psi β-barrels." *in preparation*.

Coux, O., Tanaka, K. and Goldberg, A. L. (1996). "Structure and functions of the 20S and 26S proteasomes." *Annu Rev Biochem* **65**: 801-47.

Cuff, J. A. and G. J. Barton (2000). "Application of multiple sequence alignment profiles to improve protein secondary structure prediction." *Proteins* **40**(3): 502-11.

De Mot, R., Nagy, I. and Baumeister, W. (1998). "A self-compartmentalizing protease in Rhodococcus: the 20S proteasome." *Antonie Van Leeuwenhoek* **74**(1-3): 83-7.

DeLaBarre, B. and Brunger, A. T. (2005). "Nucleotide dependent motion and mechanism of action of p97/VCP." *J Mol Biol* **347**(2): 437-52.

Djuranovic, S., Rockel, B., Lupas, A. N. and Martin, J. (2006). "Characterization of AMA, a new AAA protein from Archaeoglobus and methanogenic archaea." *J Struct Biol*. epub

Dong, T. C., Cutting, S. M. and Lewis, R. J. (2004). "DNA-binding studies on the Bacillus subtilis transcriptional regulator and AbrB homologue, SpoVT." *FEMS Microbiol Lett* **233**(2): 247-56.

Dougan, D. A., Mogk, A., Zeth, K., Turgay, K. and Bukau, B. (2002). "AAA+ proteins and substrate recognition, it all depends on their partner in crime." *FEBS Lett* **529**(1): 6-10.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic Acids Res* **32**(5): 1792-7.

Erdmann, R., Wiebel, F. F., Flessau, A., Rytka, J., Beyer, A., Frohlich, K. U. and Kunau, W. H. (1991). "PAS1, a yeast gene required for peroxisome biogenesis, encodes a member of a novel family of putative ATPases." *Cell* **64**(3): 499-510.

Finley, D., Tanaka, K., Mann, C., Feldmann, H., Hochstrasser, M., Vierstra, R., Johnston, S., Hampton, R., Haber, J., McCusker, J., Silver, P., Frontali, L., Thorsness, P., Varshavsky, A., Byers, B., Madura, K., Reed, S. I., Wolf, D., Jentsch, S., Baumeister, W., Goldberg, A., Fried, V., Rubin, D. M., Toh-e, A.. (1998). "Unified nomenclature for subunits of the Saccharomyces cerevisiae proteasome regulatory particle." *Trends Biochem Sci* **23**(7): 244-5.

Frickey, T. and Lupas, A. N. (2004). "Phylogenetic analysis of AAA proteins." *J Struct Biol* **146**(1-2): 2-10.

Frohlich, K. U. (2001). "An AAA family tree." *J Cell Sci* **114**(Pt 9): 1601-2.

Gentry, D. R., Hernandez, V. J., Nguyen, L. H., Jensen, D. B. and Cashel, M. (1993). "Synthesis of the stationary-phase sigma factor sigma s is positively regulated by ppGpp." *J Bacteriol* **175**(24): 7982-9.

Gerega, A., Rockel, B., Peters, J., Tamura, T., Baumeister, W. and Zwickl, P. (2005). "VAT, the thermoplasma homolog of mammalian p97/VCP, is an N domain regulated protein unfoldase." *J Biol Chem*.

Ghislain, M., Dohmen, R. J., Levy, F. and Varshavsky, A. (1996). "Cdc48p interacts with Ufd3p, a WD repeat protein required for ubiquitin-mediated proteolysis in Saccharomyces cerevisiae." *Embo J* **15**(18): 4884-99.

Golbik, R., Lupas, A. N., Koretke, K. K., Baumeister, W. and Peters, J. (1999). "The Janus face of the archaeal Cdc48/p97 homologue VAT: protein folding versus unfolding." *Biol Chem* **380**(9): 1049-62.

Grishin, N. V. (2001). "Fold change in evolution of protein structures." *J Struct Biol* **134**(2-3): 167-85.

Hammarstrom, M., Hellgren N., van Den Berg, S., Berglund, H. and Hard, T. (2002). "Rapid screening for improved solubility of small human proteins produced as fusion proteins in Escherichia coli." *Protein Sci* **11**(2): 313-21.

Hartman, J. J. and Vale, R. D. (1999). "Microtubule disassembly by ATP-dependent oligomerization of the AAA enzyme katanin." *Science* **286**(5440): 782-5.

Hinnerwisch, J., Reid, B. G., Fenton, W. A. and Horwich, A. L. (2005). "Roles of the N-domains of the ClpA unfoldase in binding substrate proteins and in stable complex formation with the ClpP protease." *J Biol Chem* **280**(49): 40838-44.

Jones, D. T. (1999). "Protein secondary structure prediction based on position-specific scoring matrices." *J Mol Biol* **292**(2): 195-202.

Karata, K., Inagawa, T., Wilkinson, A. J., Tatsuta, T. and Ogura, T. (1999). "Dissecting the role of a conserved motif (the second region of homology) in the AAA family of ATPases. Site-directed mutagenesis of the ATP-dependent protease FtsH." *J Biol Chem* **274**(37): 26225-32.

Katz, M. E., Strugnell, R. A. and Rood, J. I. (1992). "Molecular characterization of a genomic region associated with virulence in Dichelobacter nodosus." *Infect Immun* **60**(11): 4586-92.

Kelley, L. A., MacCallum, R. M. and Sternberg, M. J. (2000). "Enhanced genome annotation using structural profiles in the program 3D-PSSM." *J Mol Biol* **299**(2): 499-520.

Krzywda, S., Brzozowski, A. M., Verma, C., Karata, K., Ogura, T. and Wilkinson, A. J. (2002). "The crystal structure of the AAA domain of the ATP-dependent protease FtsH of Escherichia coli at 1.5 A resolution." *Structure* **10**(8): 1073-83.

Kunau, W. H., Beyer, A., Franken, T., Gotte, K., Marzioch, M., Saidowsky, J., Skaletz-Rorowski, A. and Wiebel, F. F. (1993). "Two complementary approaches to study peroxisome biogenesis in Saccharomyces cerevisiae: forward and reversed genetics." *Biochimie* **75**(3-4): 209-24.

Lanzetta, P. A., Alvarez, L. J., Reinach, P. S. and Candia, O. A. (1979). "An improved assay for nanomole amounts of inorganic phosphate." *Anal Biochem* **100**(1): 95-7.

Latterich, M., Frohlich, K. U. and Schekman, R. (1995). "Membrane fusion and the cell cycle: Cdc48p participates in the fusion of ER membranes." *Cell* **82**(6): 885-93.

Levdikov, V. M., Blagova, E., Joseph, P., Sonenshein, A. L. and Wilkinson, A. J. (2006). "The structure of CodY, a GTP- and isoleucine-responsive regulator of stationary phase and virulence in gram-positive bacteria." *J Biol Chem* **281**(16): 11366-73.

Lo, J. H., Baker, T. A. and Sauer, R. T. (2001). "Characterization of the N-terminal repeat domain of Escherichia coli ClpA-A class I Clp/HSP100 ATPase." *Protein Sci* **10**(3): 551-9.

Loris, R., Marianovsky, I., Lah, J., Laeremans, T., Engelberg-Kulka, H., Glaser, G., Muyldermans, S. and Wyns, L. (2003). "Crystal structure of the intrinsically flexible addiction antidote MazE." *J Biol Chem* **278**(30): 28252-7.

Lowe, J., Stock, D., Jap, B., Zwickl, P., Baumeister, W. and Huber, R. (1995). "Crystal

structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 A resolution." *Science* **268**(5210): 533-9.

Lupas, A. N. and Martin, J. (2002). "AAA proteins." *Curr Opin Struct Biol* **12**(6): 746-53.

Maurizi, M. R. and Li, C. C. (2001). "AAA proteins: in search of a common molecular basis. International Meeting on Cellular Functions of AAA Proteins." *EMBO Rep* **2**(11): 980-5.

McGuffin, L. J., Bryson, K. and Jones, D. T. (2000). "The PSIPRED protein structure prediction server." *Bioinformatics* **16**(4): 404-5.

Medalia, N., Sharon, M., Martinez-Arias, R., Mihalache, O., Robinson, C. V., Medalia, O. and Zwickl, P. (2006). "Functional and structural characterization of the Methanosarcina mazei proteasome and PAN complexes." *J Struct Biol*.

Meyer, H. H., Shorter, J. G., Seemann, J., Pappin, D. and Warren, G. (2000). "A complex of mammalian ufd1 and npl4 links the AAA-ATPase, p97, to ubiquitin and nuclear transport pathways." *Embo J* **19**(10): 2181-92.

Mogk, A., Schlieker, C., Strub, C., Rist, W., Weibezahn, J. and Bukau, B. (2003). "Roles of individual domains and conserved motifs of the AAA+ chaperone ClpB in oligomerization, ATP hydrolysis, and chaperone activity." *J Biol Chem* **278**(20): 17615-24.

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." *J Mol Biol* **247**(4): 536-40.

Neuwald, A. F., Aravind, L., Spouge, J. L. and Koonin, E. V. (1999). "AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes." *Genome Res* **9**(1): 27-43.

Niwa, H., Tsuchiya, D., Makyio, H., Yoshida, M. and Morikawa, K. (2002). "Hexameric ring structure of the ATPase domain of the membrane-integrated metalloprotease FtsH from Thermus thermophilus HB8." *Structure* **10**(10): 1415-23.

Ogura, T., Whiteheart, S. W. and Wilkinson, A. J. (2004). "Conserved arginine residues implicated in ATP hydrolysis, nucleotide-sensing, and inter-subunit interactions in AAA and AAA+ ATPases." *J Struct Biol* **146**(1-2): 106-12.

Ogura, T. and Wilkinson, A. J. (2001). "AAA+ superfamily ATPases: common structure--diverse function." *Genes Cells* **6**(7): 575-97.

Pamnani, V., Tamura, T., Lupas, A., Peters, J., Cejka, Z., Ashraf, W. and Baumeister, W. (1997). "Cloning, sequencing and expression of VAT, a CDC48/p97 ATPase homologue from the archaeon Thermoplasma acidophilum." *FEBS Lett* **404**(2-3): 263-8.

Park, S., Isaacson, R., Kim, H. T., Silver, P. A. and Wagner, G. (2005). "Ufd1 exhibits the AAA-ATPase fold with two distinct ubiquitin interaction sites." *Structure* **13**(7): 995-1005.

Peters, J. M., Walsh, M. J. and Franke, W. W. (1990). "An abundant and ubiquitous homo-oligomeric ring-shaped ATPase particle related to the putative vesicle fusion proteins Sec18p and NSF." *Embo J* **9**(6): 1757-67.

Ratnayake-Lecamwasam, M., Serror, P., Wong, K. W. and Sonenshein, A. L. (2001). "Bacillus subtilis CodY represses early-stationary-phase genes by sensing GTP levels." *Genes Dev* **15**(9): 1093-103.

Reuter, C. J., Kaczowka, S. J. and Maupin-Furlow, J. A. (2004). "Differential regulation of the PanA and PanB proteasome-activating nucleotidase and 20S proteasomal proteins of the haloarchaeon Haloferax volcanii." *J Bacteriol* **186**(22): 7763-72.

Richter, G., Krieger, C., Volk, R., Kis, K., Ritz, H., Gotze, E. and Bacher, A. (1997). "Biosynthesis of riboflavin: 3,4-dihydroxy-2-butanone-4-phosphate synthase." *Methods Enzymol* **280**: 374-82.

Rockel, B., Jakana, J., Chiu, W. and Baumeister, W. (2002). "Electron cryo-microscopy of VAT, the archaeal p97/CDC48 homologue from Thermoplasma acidophilum." *J Mol Biol* **317**(5): 673-81.

Rockel, B., Walz, J., Hegerl, R., Peters, J., Typke, D. and Baumeister, W. (1999). "Structure of VAT, a CDC48/p97 ATPase homologue from the archaeon Thermoplasma acidophilum as studied by electron tomography." *FEBS Lett* **451**(1): 27-32.

Ruepp, A., Rockel, B., Gutsche, I., Baumeister, W. and Lupas, A. N. (2001). "The Chaperones of the archaeon Thermoplasma acidophilum." *J Struct Biol* **135**(2): 126-38.

Schuler, G. D., S. F. Altschul, et al. (1991). "A workbench for multiple alignment construction and analysis." *Proteins* **9**(3): 180-90.

Schultz, S. C., Shields, G. C. and Steitz, T. A. (1991). "Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees." *Science* **253**(5023): 1001-7.

Scott, A., Chung, H. Y., Gonciarz-Swiatek, M., Hill, G. C., Whitby, F. G., Gaspar, J., Holton, J. M., Viswanathan, R., Ghaffarian, S., Hill, C. P. and Sundquist, W. I. (2005). "Structural and mechanistic studies of VPS4 proteins." *Embo J* **24**(20): 3658-69.

Siddiqui, S. M., Sauer, R. T. and Baker, T. A. (2004). "Role of the processing pore of the ClpX AAA+ ATPase in the recognition and engagement of specific protein substrates." *Genes Dev* **18**(4): 369-74.

Singh, S. K., Guo, F. and Maurizi, M. R. (1999). "ClpA and ClpP remain associated during multiple rounds of ATP-dependent protein degradation by ClpAP protease." *Biochemistry* **38**(45): 14906-15.

Soding, J., Biegert, A. and Lupas, A. N. (2005). "The HHpred interactive server for protein homology detection and structure prediction." *Nucleic Acids Res* **33**(Web Server issue): W244-8.

Strickland, E., Hakala, K., Thomas, P. J. and DeMartino, G. N. (2000). "Recognition of misfolding proteins by PA700, the regulatory subcomplex of the 26 S proteasome." *J Biol Chem* **275**(8): 5565-72.

Suno, R., Niwa, H., Tsuchiya, D., Zhang, X., Yoshida, M. and Morikawa, K. (2006). "Structure of the whole cytosolic region of ATP-dependent protease FtsH." *Mol Cell* **22**(5): 575-85.

Swaffield, J. C., Bromberg, J. F. and Johnston, S. A. (1992). "Alterations in a yeast protein resembling HIV Tat-binding protein relieve requirement for an acidic activation domain in GAL4." *Nature* **360**(6406): 768.

Thayer, M. M., Ahern, H., Xing, D., Cunningham, R. P. and Tainer, J. A. (1995). "Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure." *Embo J* **14**(16): 4108-20.

Tomoyasu, T., Yamanaka, K., Murata, K., Suzaki, T., Bouloc, P., Kato, A., Niki, H., Hiraga, S. and Ogura, T. (1993). "Topology and subcellular localization of FtsH protein in Escherichia coli." *J Bacteriol* **175**(5): 1352-7.

Vale, R. D. (2000). "AAA proteins. Lords of the ring." *J Cell Biol* **150**(1): F13-9.

Vaughn, J. L., Feher, V., Naylor, S., Strauch, M. A. and Cavanagh, J. (2000). "Novel DNA binding domain and genetic regulation model of Bacillus subtilis transition state regulator abrB." *Nat Struct Biol* **7**(12): 1139-46.

Vicente, M., Gomez, M. J. and Ayala, J. A. (1998). "Regulation of transcription of cell division genes in the Escherichia coli dcw cluster." *Cell Mol Life Sci* **54**(4): 317-24.

Voges, D., Zwickl, P. and Baumeister, W. (1999). "The 26S proteasome: a molecular machine designed for controlled proteolysis." *Annu Rev Biochem* **68**: 1015-68.

Walker, J. E., Saraste, M., Runswick, M. J. and Gay, N. J. (1982). "Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold." *Embo J* **1**(8): 945-51.

Wanner, B. L. (1993). "Gene regulation by phosphate in enteric bacteria." *J Cell Biochem* **51**(1): 47-54.

Wilson, H. L., Ou, M. S., Aldrich, H. C. and Maupin-Furlow, J. (2000). "Biochemical and physical properties of the Methanococcus jannaschii 20S proteasome and PAN, a homolog of the ATPase (Rpt) subunits of the eucaryal 26S proteasome." *J Bacteriol* **182**(6): 1680-92.

Wolf, S., Nagy, I., Lupas, A., Pfeifer, G., Cejka, Z., Muller, S. A., Engel, A., De Mot, R. and Baumeister, W. (1998). "Characterization of ARC, a divergent member of the AAA ATPase family from Rhodococcus erythropolis." *J Mol Biol* **277**(1): 13-25.

Yamada, T., Okuhara, K., Iwamatsu, A., Seo, H., Ohta, K., Shibata, T. and Murofushi, H. (2000). "p97 ATPase, an ATPase involved in membrane fusion, interacts with DNA unwinding factor (DUF) that functions in DNA replication." *FEBS Lett* **466**(2-3): 287-91.

Yamada-Inagawa, T., Okuno, T., Karata, K., Yamanaka, K. and Ogura, T. (2003). "Conserved pore residues in the AAA protease FtsH are important for proteolysis and its coupling to ATP hydrolysis." *J Biol Chem* **278**(50): 50182-7.

Zhang, X., Stoffels, K., Wurzbacher, S., Schoofs, G., Pfeifer, G., Banerjee, T., Parret, A. H., Baumeister, W., De Mot, R. and Zwickl, P. (2004). "The N-terminal coiled coil of the Rhodococcus erythropolis ARC AAA ATPase is neither necessary for oligomerization nor nucleotide hydrolysis." *J Struct Biol* **146**(1-2): 155-65.

Zhang, Y., Zhang, J., Hoeflich, K. P., Ikura, M., Qing, G. and Inouye, M. (2003). "MazF cleaves cellular mRNAs specifically at ACA to block protein synthesis in Escherichia coli." *Mol Cell* **12**(4): 913-23.

Zwickl, P. (2002). "The 20S proteasome." *Curr Top Microbiol Immunol* **268**: 23-41.

Zwickl, P., Ng, D., Woo, K. M., Klenk, H. P. and Goldberg, A. L. (1999). "An archaebacterial ATPase, homologous to ATPases in the eukaryotic 26 S proteasome, activates protein breakdown by 20 S proteasomes." *J Biol Chem* **274**(37): 26008-14.

Meine akademischen Lehrer neben Prof. Dr. G. Dodt und Prof. Dr. A.N. Lupas waren die Damen und Herren Professoren und Doktoren:

Vuckovic, G.	Anorganische Chemie I
Pavlovic, V.	Organische Chemie I/ Organische Chemie II
Tesic, Z.	Analytische Chemie I
Bukvic, S.	Physik
Vrecica, S.	Mathematik
Popovic, O.	Englisch I/ Englisch II
Serban, N.	Biologie
Glintic, T.	Wissenschaftsphilosophie
Konjevic, R	Physikalische Chemie
Niketic, S.	Anorganische Chemie II
Micovic, Lj.	Organische Chemie III
Jankov, R.M.	Naturstoff-Chemie/Immunochemie
Niketic, V.	Biochemie I
Todorovic, M.	Analytisch Chemie II
Vajs, V.	Strukturaufkärung
Matic, G.	Biochemie II/ Molekularendokrinologie
Petrovic, Dj.	Enzymatik I
Spasic, M.	Biochemie III
Vucetic, J.	Mikrobiologische Chemie/Biotechnologie
Knezevic, J.	Mikrobiologie
Soskic, V.	Experimentalbiochemie
Romac, S.	Molekularbiologie
Tomic, M	Enzymatik II
Davidovic, V.	Physiologie
Konjovic, R.	Pflanzenphysiologie
Kataranovska, M.	Immunobiologie
Nedic, O.	Pathobiochemie
Vasiljevic, B.	Moleculare Genetik/Molecularbiologie von Krebszellen
Milojkovic-Opsenica, D.	Statistik
Duszenko, M.	Biochemie

Zur Person		Sergej Djuranovic Geboren am 6. August 1975 in Knin Kroatien
Schulausbildung	9/1982-6/1990 9/1990-6/1994	Grundschule Slobodan Skerovic, Gymnasium, Podgorica, Montenegro
Studium	10/1994- 22.07/1999	Studium der Biochemie an der Fakultät für Chemie der Universität Belgrad in Serbien und Montenegro
	10/1999-12/2001	Magisterarbeit in Neurobiochemie an der Fakultät für Chemie der Universität Belgrad in Serbien und Montenegro
Praktische Ausbildung	11/2001-07/2002	gearbeitet in der Kinderklinik in Podgorica, Serbien und Montenegro
Promotion	10/2002 -03/2006	Durchführung des experimentellen Teils der vorliegenden Arbeit „Evolution der Substraterkennungsdomänen von AAA Proteinen" am Max-Planck-Institut für Entwicklungsbiologie-Tübingen unter der Anleitung von Prof Dr. Andrei Lupas und Prof Dr. Gabriele Dodt

ZUSAMMENFASSUNG

AAA Proteine gehören zur großen Überfamilie der AAA+ Proteine. Diese sind ringförmige P-loop NTPasen, deren gemeinsame Funktion das engerieabhängige Entfalten von Makromolekülen ist. AAA Proteine bestehen im Allgemeinen aus einer N-terminalen Domäne und einer oder zwei ATPase Domänen, die mit D1 und D2 bezeichnet werden. Die ATPase Domänen innerhalb der Familie der AAA Proteine sind verhältnismäßig konserviert. Sie bewirken wahrscheinlich eine Hexamerisierung. Die N-terminalen Domänen sind wichtig für die Erkennung und Bindung des Substrats und unterscheiden sich im Gegensatz zu den ATPase Domänen strukturell.

Auf der Grundlage veröffentlichter Daten und zusätzlicher bioinformatischer Analyse der AAA Proteine wurden mehrere verschiedene N-terminale Domänen von AAA Proteinen aus Archaebakterien für eine funktionelle und strukturelle Charakterisierung ausgewählt. Es wurden weiterhin Proteine charakterisiert, die ähnliche oder verwandte Domänen wie die AAA Proteine haben. Dabei wurde darauf Wert gelegt, Zusammenhänge und Gemeinsamkeiten herauszustellen. Es wurden Hitze- und chemische Aggregations Tests mit verschiedenen Substratproteinen durchgeführt, um die N-terminalen Domänen oder die gesamten AAA Proteine auf intrinsische Chaperon-Aktivität zu untersuchen. Die Proteinstrukturen wurden mit Röntgenstrukturanalyse oder mit NMR Spektroskopie bestimmt. Die Ergebnisse dieser Arbeit zeigen, dass die barrel-fömigen N-terminalen Domänen der AAA Proteine aus bauart-ähnlichen, Nukleinsäure bindenden Domänen hervorgingen. Die Substrataffinität hat sich wahrscheinlich durch verschiedene evolutionäre Mechanismen von Nukleinsäuren zu Proteinen gewandelt. Im Fall der doppelten Psi-Barrels ist dies vermutlich durch die Evolution eines einfachen βαβ-Motifs geschehen, das man auch in den RIFT und in den swapped hairpin-Barrels findet. Letztere sind Transkriptionsfaktoren, d.h. sie binden DNA. Die Struktur von Mj0056 und SpoVT legt nahe, dass RIFT und swapped hairpin-Barrels entweder durch eine Insertion verschiedener Strukturelemente (Mj0056) oder durch die Rekrutierung einer Domäne (SpoVT) weiter evolviert sind.

Ähnlichkeiten zwischen den N-terminalen Domänen von PAN und ARC wurden sowohl in Struktur und Funktion gefunden. Beide Domänen beinhalten eine coiled-coil Faltung, auf die eine oder zwei OB Faltungen folgen. OB steht für Oligosaccharid-Bindung und weist darauf hin, dass diese Domäne ebenfalls aus einer DNA-bindenden Struktur entstand. Die Struktur der ARC-Nc Subdomäne und eine umfassende Analyse von Chimären aus coiled-coil und OB Faltungen zeigen auf, dass sich die N-terminalen Domänen von ARC und PAN evolutionär durch die Rekrutierung von Domänen entwickelt haben. Die strikte strukturelle Zusammensetzung der Subdomänen ist wichtig für die Chaperon-Funktion und wird durch die konservierten PP-Linker, die die beiden Subdomänen verbinden, erhalten.

Die strukturelle und funktionale Charakterisierung von AfAMA, das zu der neuen Gruppe der AMA AAA Proteine gehört, zeigte, dass die Substratbindefunktion und die Chaperon-Aktivität dieser Proteine in ihrer β-clam ähnlichen N-terminalen Domäne liegen. Diese Domäne kann die genannten Funktionen unabhängig voneinander erfüllen, was sie von anderen homologen Domänen unterscheidet. Es war uns möglich, die Bedeutung der Oligomerisierung für die Aktivität dieser Domänen zu zeigen. Des Weiteren haben wir gezeigt, dass die Oligomerisierung durch ein kleines GYPL-Motif vermittelt wird, das auf einem Loop liegt, der vermutlich in das Zentrum des Hexamers zeigt. Die N-terminale Domäne von AMA bewirkt die Hexamerisierung des ganzen Proteins unabhängig von dem AAA Teil des Proteins und unabhängig von ATP Verbrauch. Darin weicht sie stark von anderen Familien der AAA Proteinen ab. Funktionelle Daten von anderen β-clam Domänen (z.B. VAT-Nc, Ph1500N und Hm-clam) würden darauf hinweisen, dass diese Domänen universelle Protein-Protein Wechselwirkungsmodule darstellen.